Introduction to Medical Bioinformatics
Sep 18, 2018
Faculty of Medicine Siriraj Hospital
Mahidol University

SIRE 503: Introduction to Medical Bioinformatics

Bhoom Suktitipat, MD, PhD
[bhoom.suk@mahidol.edu]

Graduate Program in Medical Bioinformatics
Department of Biochemistry
Faculty of Medicine Siriraj Hospital
&
Integrative Computational BioScience Center
Mahidol University

Outline

- Course administration/overview
- Genetics & Genomics vs Bioinformatics
- ***The Central Dogma*** (The Genetic Basis of Life)
- Gene Structure
- Biological Data
- Example Studies
  - Finding gene(s) for complex diseases
  - Finding the gene in Mendelian disorders

Course Schedule: Every Tuesday

COURSE SCHEDULE

DATE	TIME	ROOM	Topic	FACULTY STAFF
T Sep 18	9 - 12	L1	- Overview of the course - Introduction to Medical Bioinformatics & The Central Dogma	Bhoom Suktitipat
	13 - 16	L2	The RNA World & Systems Biology	Varodom Charoensawan
T Sep 25	9 - 12	L3	Proteomics & Application of Mass Spectrometry	Kessiri Kongmanas
	13 - 16	L4	Molecular Biology Techniques & Bioinformatics Data	Wanna Thongnoppakhun
T Oct 2	9 - 12	L5	BLAST, Primer3, and Sequence Alignment	Prapat Suriyaphol
	13 - 16	L6	Next-generation sequencing technologies	Bhoom Suktitipat
T Oct 9	9 - 12	L7	Metabolomics & its application	Sakda Khumroong
	13 - 16	L8	Genome assembly & Reference Mapping	Harald Grove
T Oct 16	13 - 16	L9	Variant Calling & Annotation	Harald Grove
T Oct 30	9 - 12	L10	Overview of Machine Learning	Faculty
	13 - 16	L11	Data mining & Text mining	Apirak Hoonlor
T Nov 6	9 - 12	L12	Applying genomic data into a real clinical practice: A dawn of the 'real' precision medicine	Vip Viprakasit
	13 - 16	L13	Genetic Risk Profile	Bhoom Suktitipat
T Nov 13	9 - 12	L14	Molecular Evolution I	Pravech Ajawatanawong
	13 - 16	L15	Molecular Evolution II	Pravech Ajawatanawong
T Dec 4	9 - 12		EXAMINATION	Faculty

Grading

- Attendance: 10%
- Homework: 40%
- Term paper: 30%
- Final Exam: 20%
  - MCQ & MEQ

	Score
A	> 90
B+	[80 - 90)
B	[70 - 80)
C	< 70

Term Paper — Due Nov 15

Pick an original article from this list that is **related to bioinformatics data analysis.** Focus more on articles that use high-throughput data in their research, either genomics, transcriptomics, proteomics, metabolomics, or other big data related to biological questions and experiments. Stay away from papers that focus mainly on classical epidemiology, conventional risk factors, or biostatistics, that do not have much utilization of biological data.

1. Science
2. Nature
3. Nature Genetics
4. Nature Medicine
5. Nature Cell Biology
6. New England Journal of Medicine
7. British Medical Journal (BMJ)
8. American Journal of Human Genetics
9. Cell
10. PNAS
11. PLOS Genetics
12. PLOS Medicine
13. Genome Research
14. Genome Biology
15. Human Mutation
16. Human Molecular Genetics

Have other ground breaking papers? Let's discuss!

Term Paper

Answer the following questions (3 sections)

Section 1: What is already known on this topic

In less than three single-sentence bullet points, please summarize the state of scientific knowledge on this topic. Emphasize on **“why”** this study needed to be done.

Section 2: What this study adds

In one or two single sentence bullet points, give a simple answer to the questions **“What do we now know as a result of this study that we did not know before?”** “**Is there any implications for practice, research, policy, or public health?”**

Be brief, succinct, specific, and accurate.

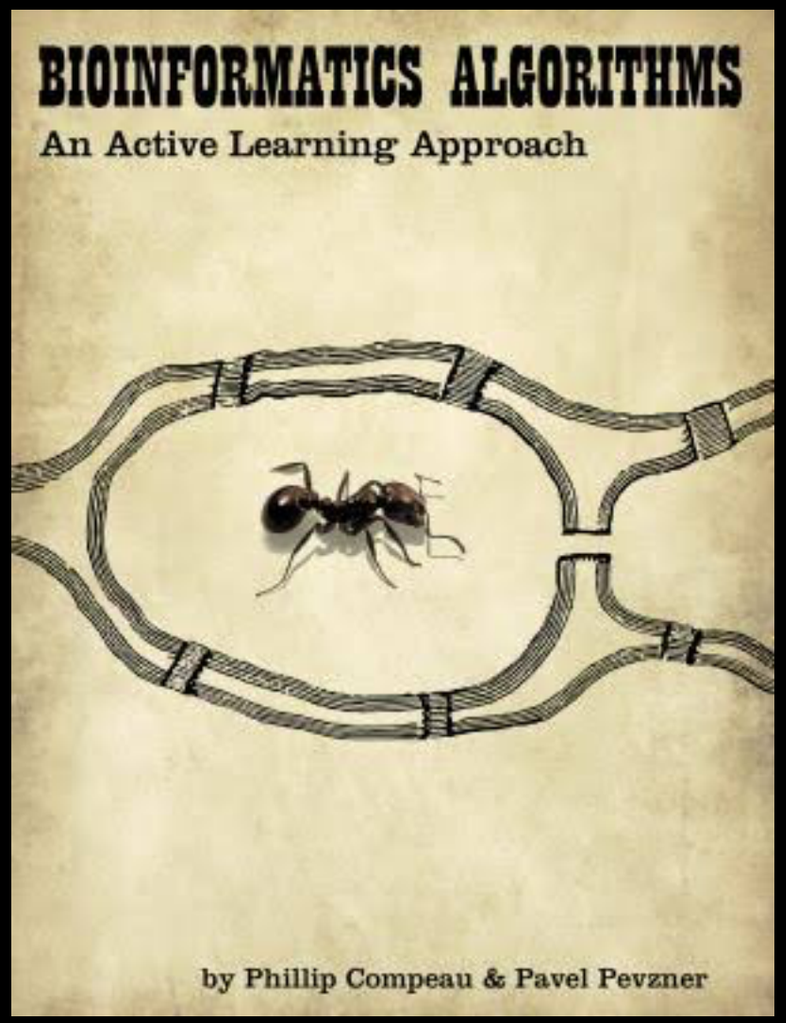
Term Paper

Section 3: How the data were analyzed

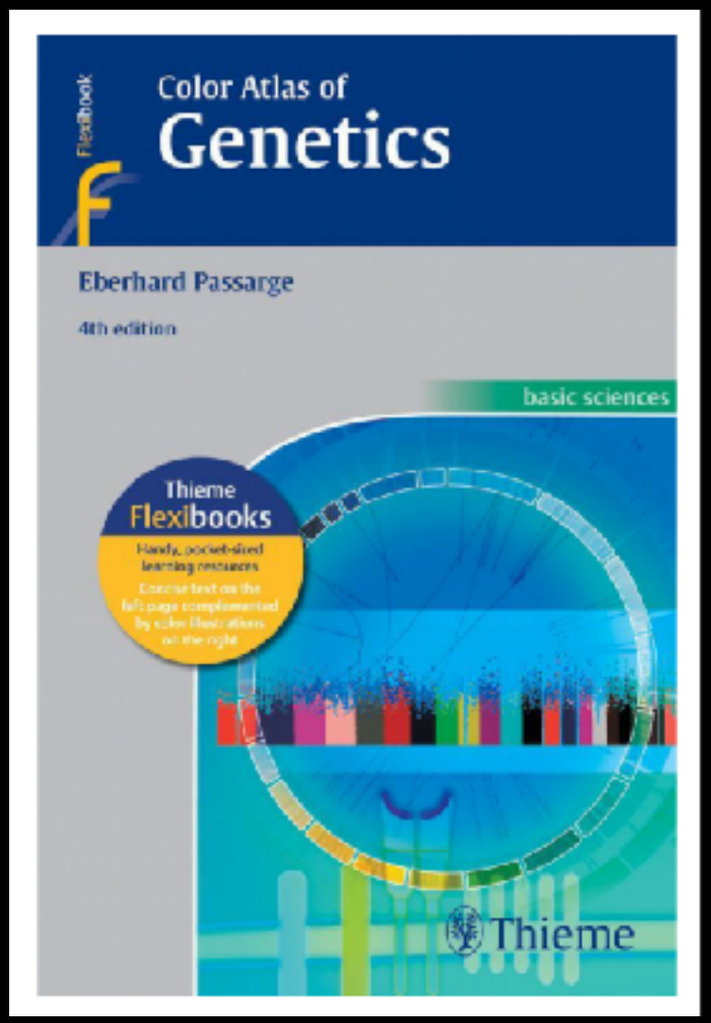
Pick 3 figures and describe the followings

1. What is question that this figure tried to answer?
2. What data have been generated to answer the question?
3. What analysis have been done to get to the conclusion?

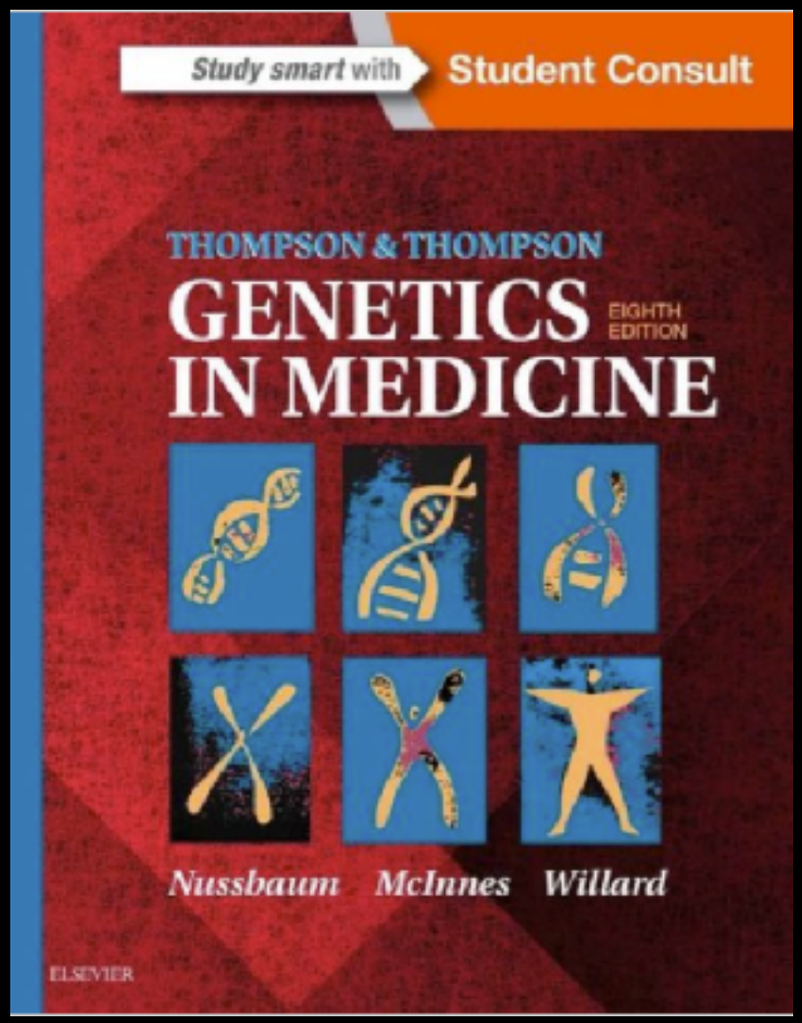
Suggested Reference



Bioinformatics Algorithms: An Active Learning Approach
Textbook by Pavel A. Pevzner and Phillip Compeau



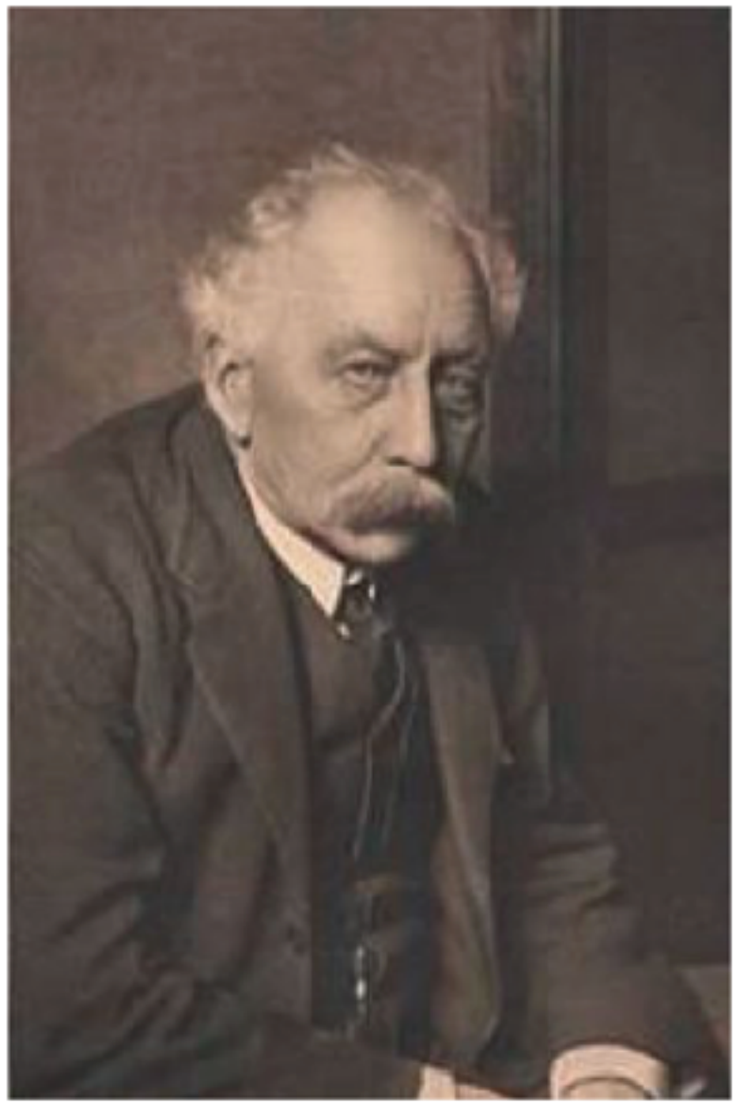
The Color Atlas of Genetics, 4th ed, 2013



Thompson & Thompson Genetics in Medicine ed.8 (2016)

Genetics and Genomics

- William Bateson
- “Genetics” —1906
- Investigate heredity and variation



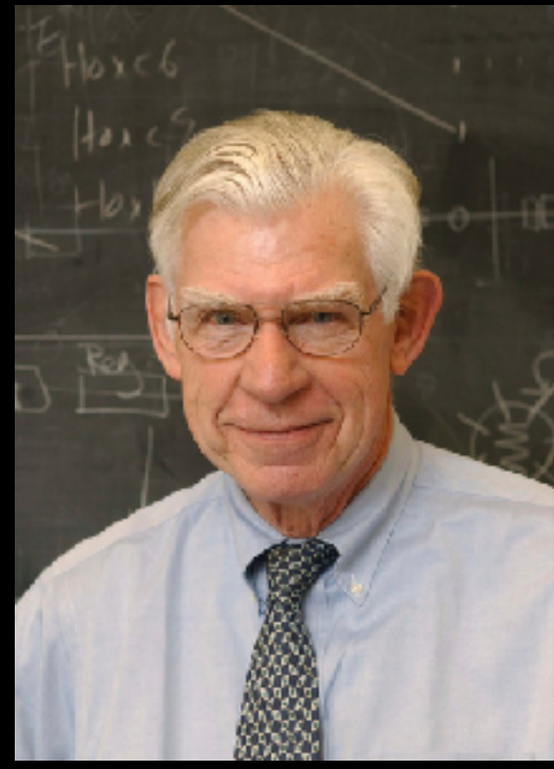
William Bateson (1861–1926)

Genetics and Genomics

- Victor A. McKusick & Francis H. Ruddle: “Genomics” 1987

  - Studying of the entire genomes

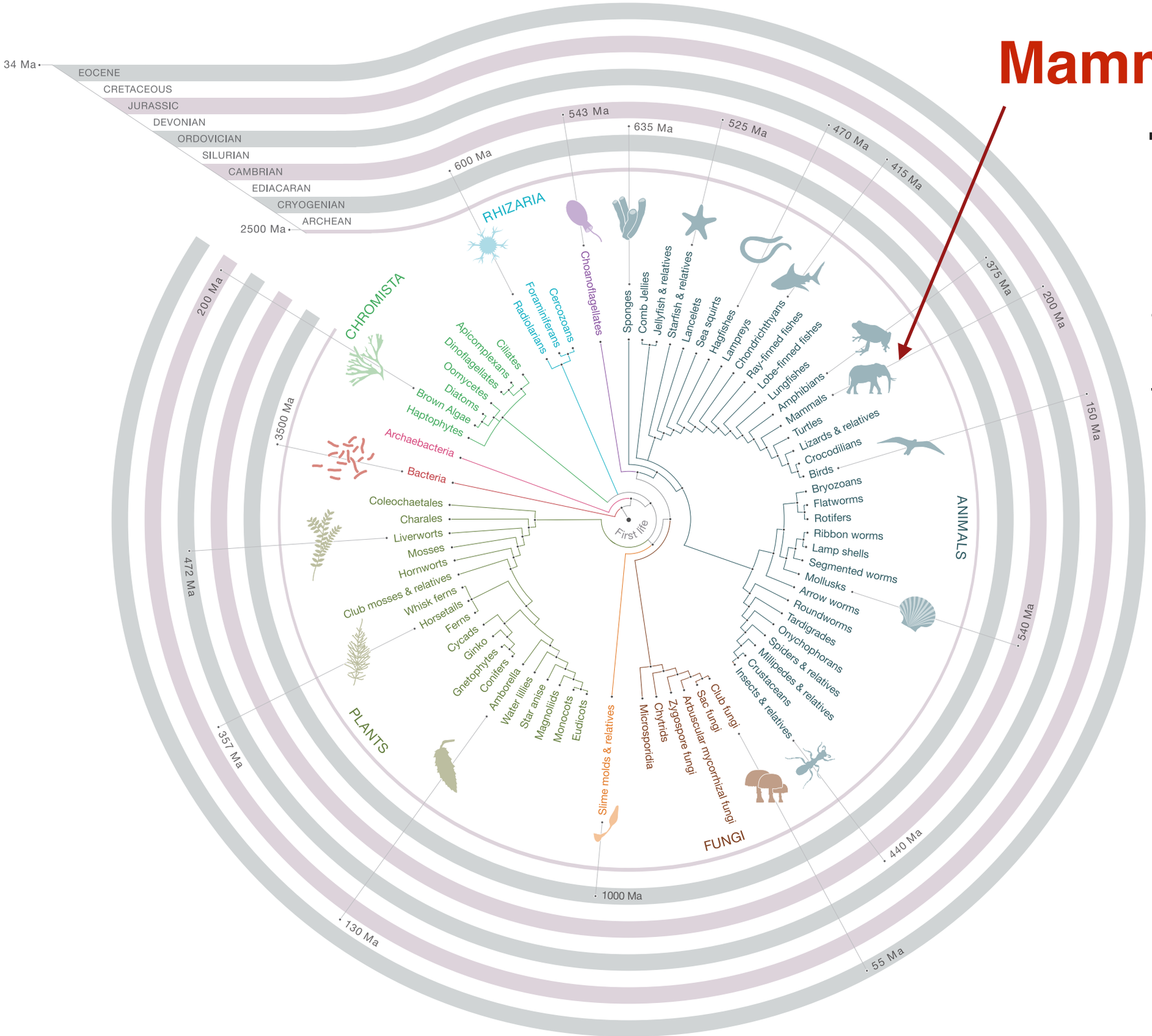
Victor McKusick
1921-2008



Francis Ruddle
1929-2013

Where we come from

Mammals

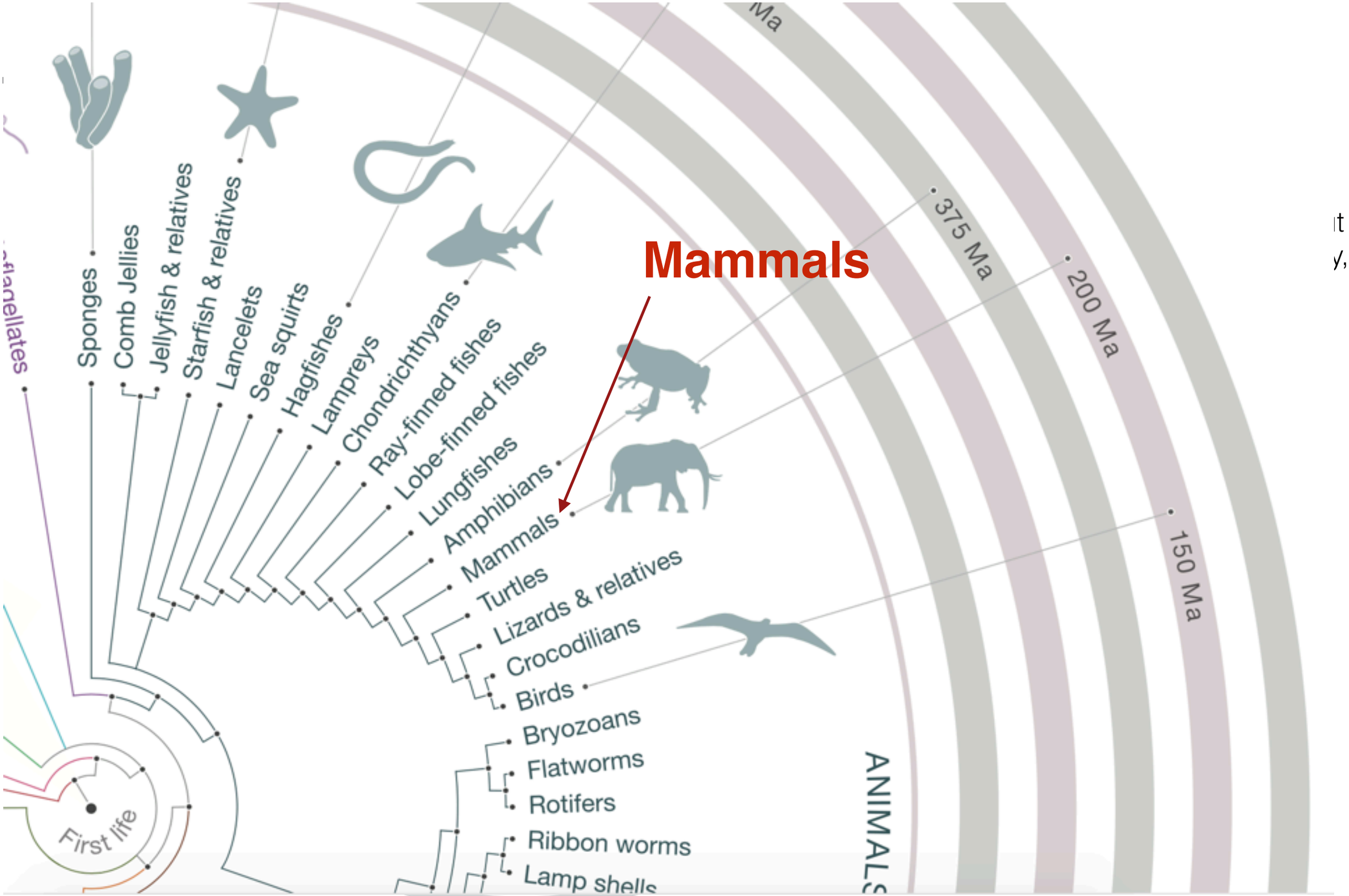


The Tree of Life

Based on research carried out by David Hillis, Derrick Zwicky, and Robin Gutell from the University of Texas
- analysis of rRNA sequences

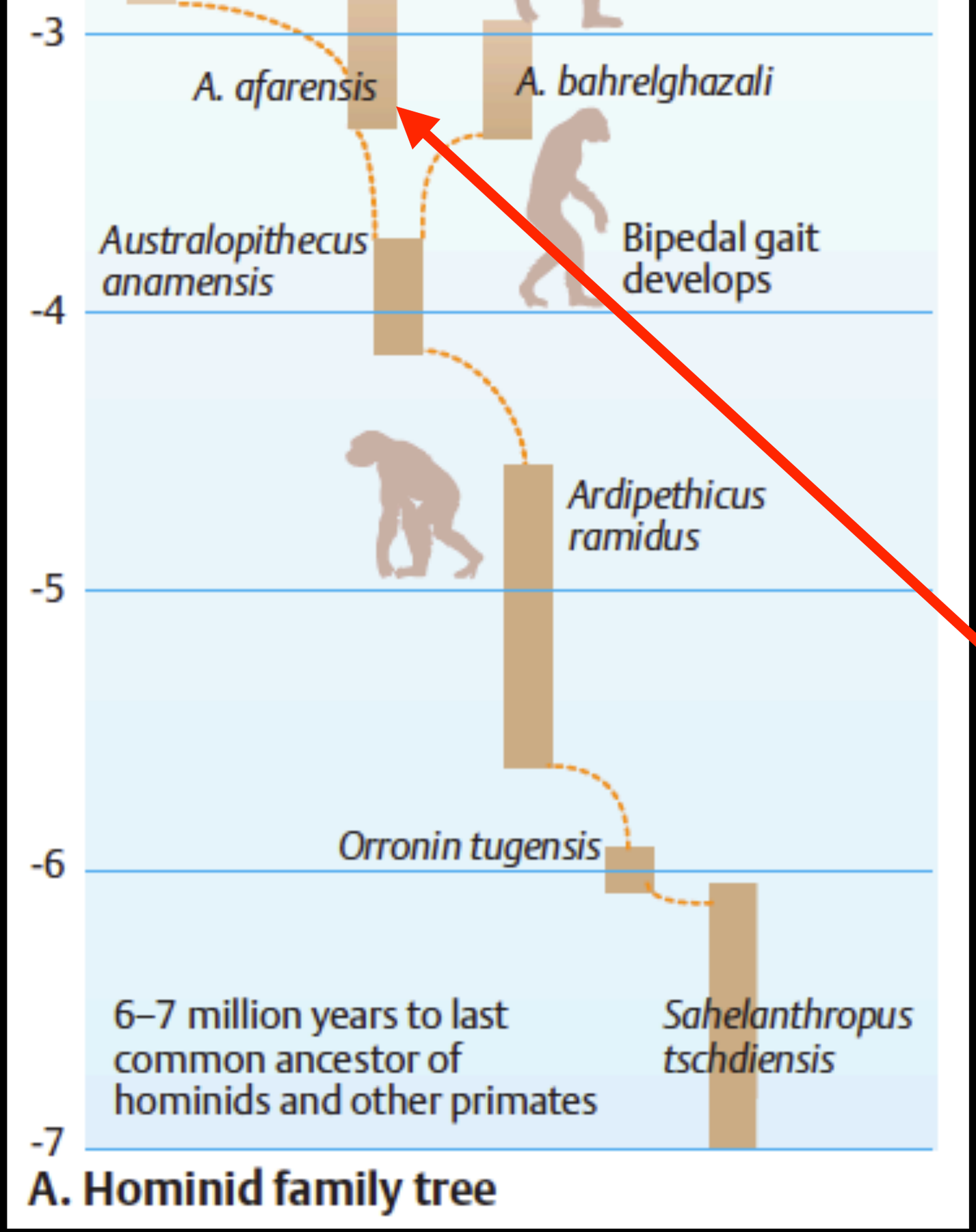
http://downloads.bbc.co.uk/nature/posters/tree_of_life.pdf

Where we come from



http://downloads.bbc.co.uk/nature/posters/tree_of_life.pdf

The Hominid Origin

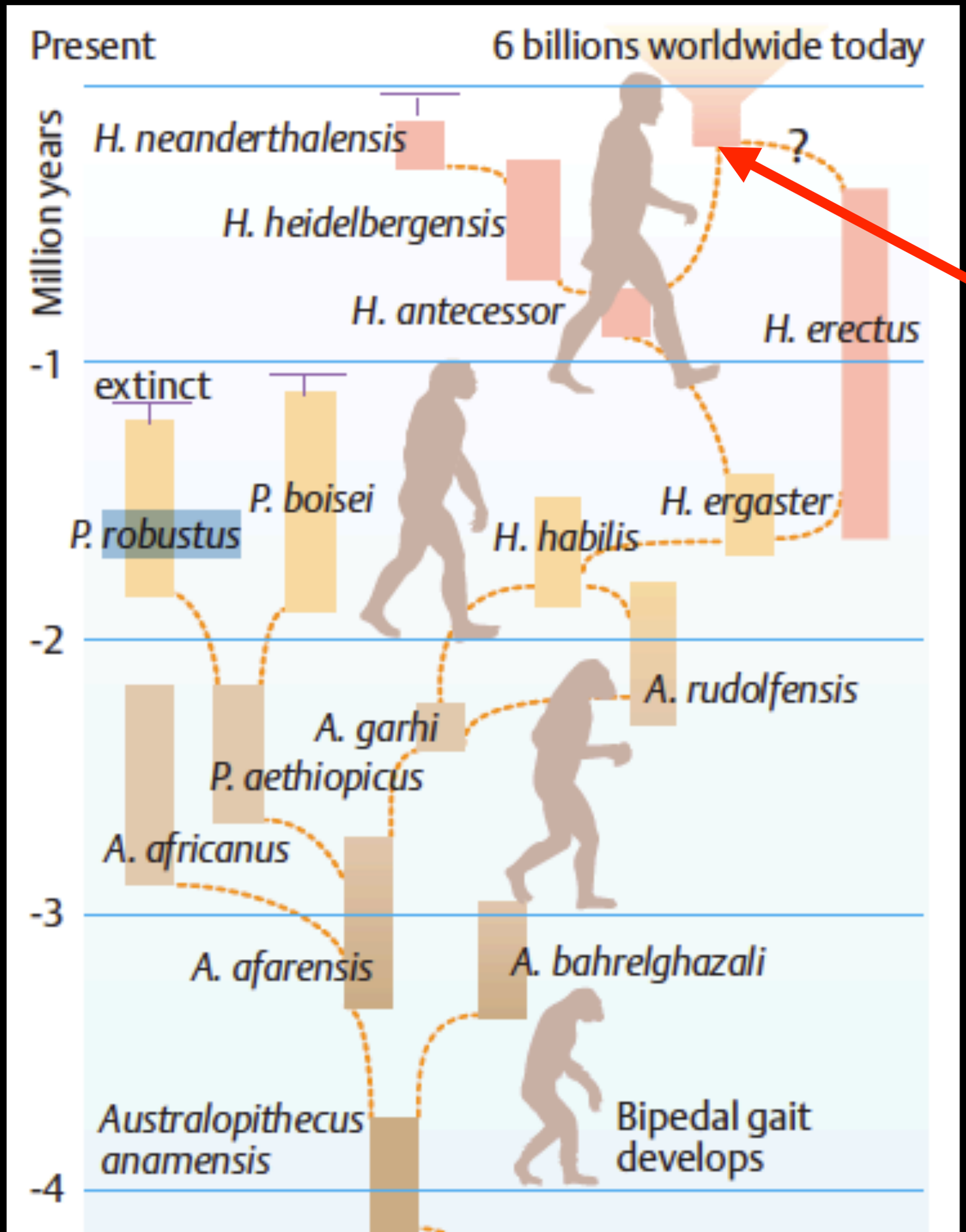


A. Hominid family tree

Lucy

Last common ancestry between human and chimps

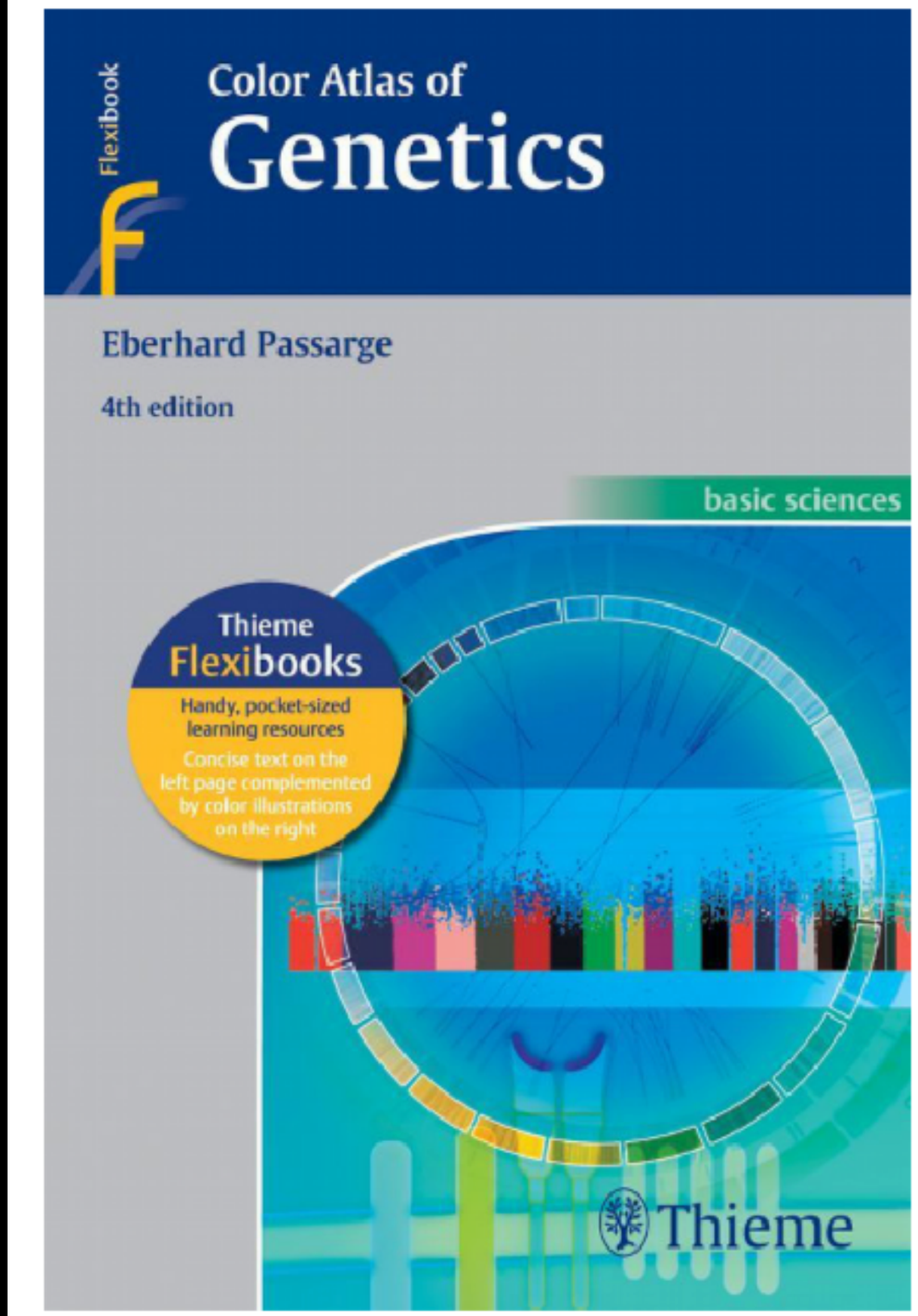
- Found in Chad (Central Africa) & Kenya (East Africa)



The Hominid Origin

Modern Human

30-40kya



The Color Atlas of Genetics, 4th ed, 2013

Why do you want to study Bioinformatics?

Biomedical Informatics?

Biomedical Informatics is, "the field that is concerned with the optimal use of information, often aided by the use of technology and people, to improve individual health, health care, public health, and biomedical research".

http://medicine.osu.edu/bmi/Pages/index.aspx

Biomedical Informatics?

The study of information and computation in biology and health. Its researchers study and manage information, study behavior related to decisions, and develop computational methods and use them to generate knowledge.

https://www.dbmi.columbia.edu/research/research-areas/

Research in Biomedical Informatics

Research areas in Biomedical Informatics

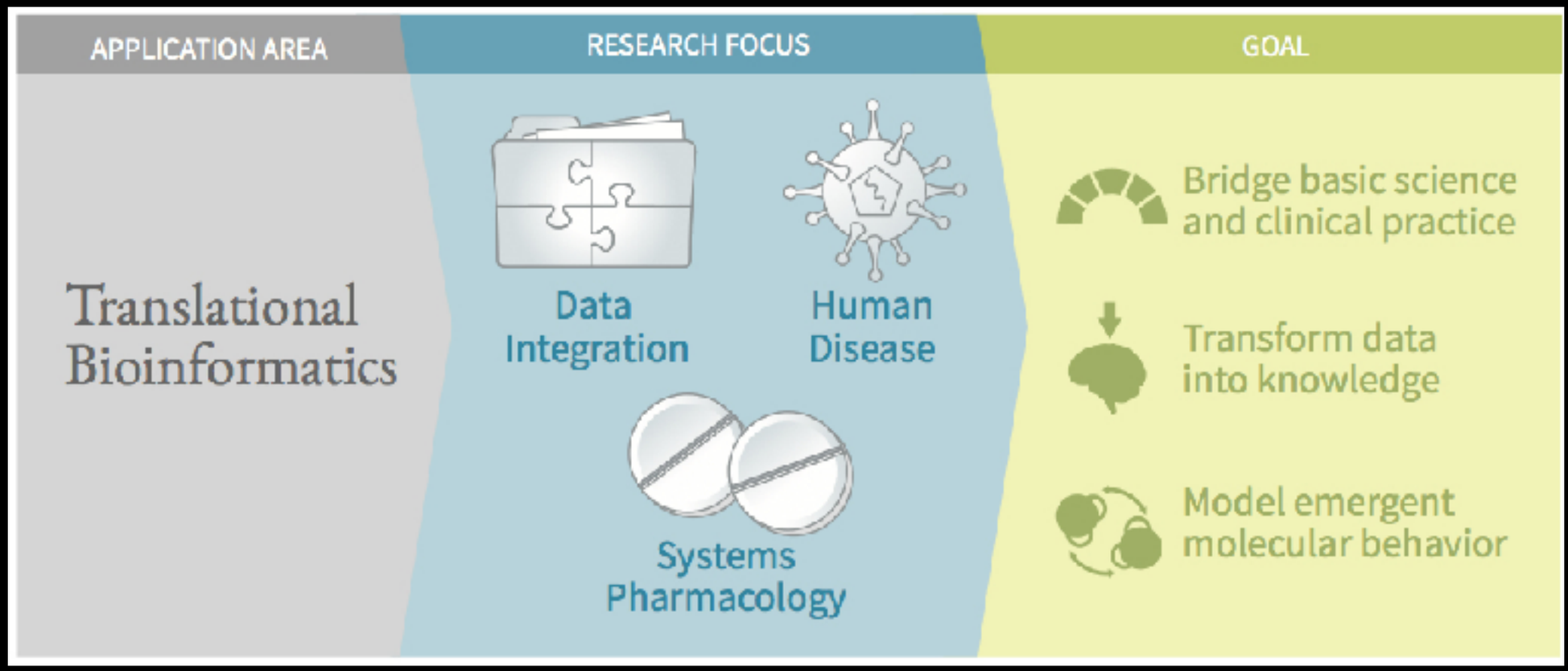
1. Clinical Informatics
2. Public Health Informatics
3. Clinical Research Informatics
4. Translational Bioinformatics
5. Computational Biology

https://www.dbmi.columbia.edu/research/research-areas/computational-biology/

Research in Biomedical Informatics

Translational Bioinformatics

Specialists in this category use their expertise to translate research-based biological discoveries into clinical utility, developing high-throughput methods to bridge research and health care.



Translational Bioinformatics

Translational Bioinformatics

The field includes:

- *Using computational and statistical analysis* to bridge the gap between basic science and clinical practice
- The *application of bioinformatics and computational biology methods* to clinical data
- *Developing models* to translate knowledge generated in model systems (mouse, fly, yeast) into knowledge about human disease
- Using data gathered on patients and human disease and *relating it back to basic science principles*

Bioinformatics

Bioinformatics /ˌbaɪ.oʊˌɪnfərˈmætɪks/ is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, **bioinformatics** combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data.

Bioinformatics - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Bioinformatics** Wikipedia

More about Bioinformatics

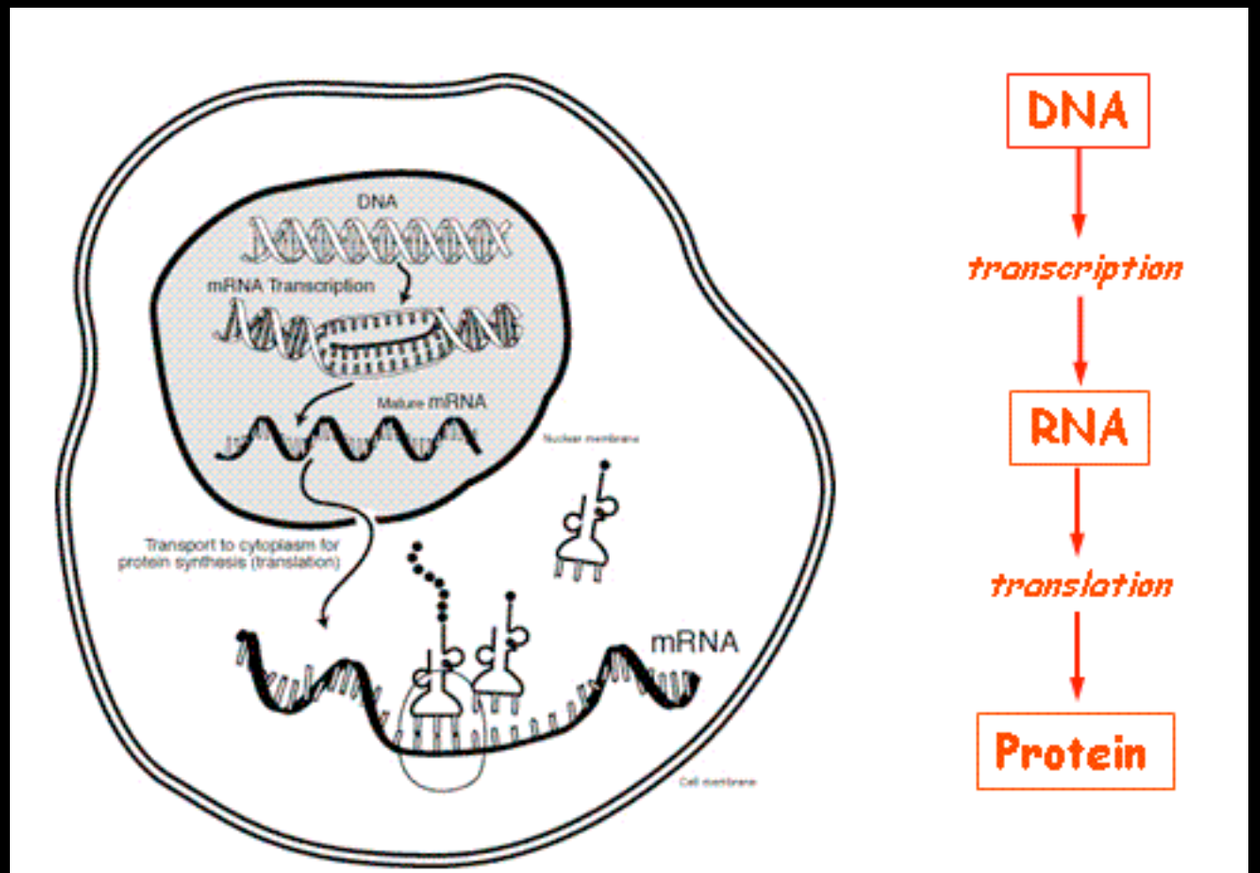
Topics in Bioinformatics

- **"Biological data resources"**
- **"Nucleic acid analysis"**
- **"Protein analysis"**
- "Sequence analysis"
- "Structure analysis"
- **"Phylogenetics"**
- **"Proteomics"**
- "Data handling"
- "Chemoinformatics"
- **"Transcriptomics"**
- "Literature and reference"
- "Ontologies, nomenclature and classification"
- **"Genetics"**
- **"Systems biology"**
- "Ecoinformatics"
- "**Genomics**"
- "Immunoinformatics"

src: http://edamontology.org/page

Bioinformatics Data?

- Biological data ~ Biomarker data
  - DNA
  - RNA
  - Protein

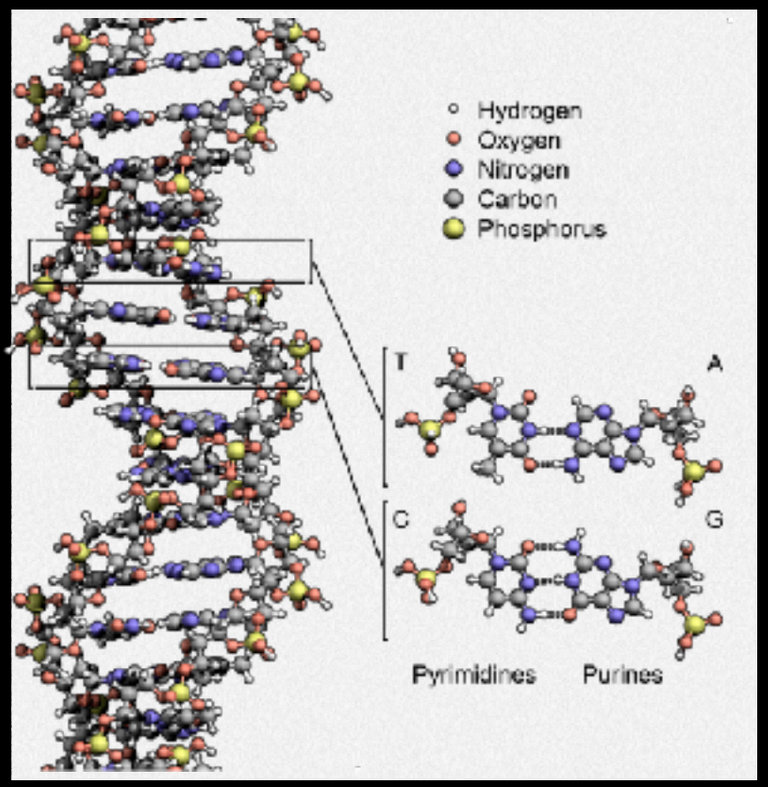


Central Dogma

Src: http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/central_dogma.html

The Central Dogma

The Genetic Basis of Life



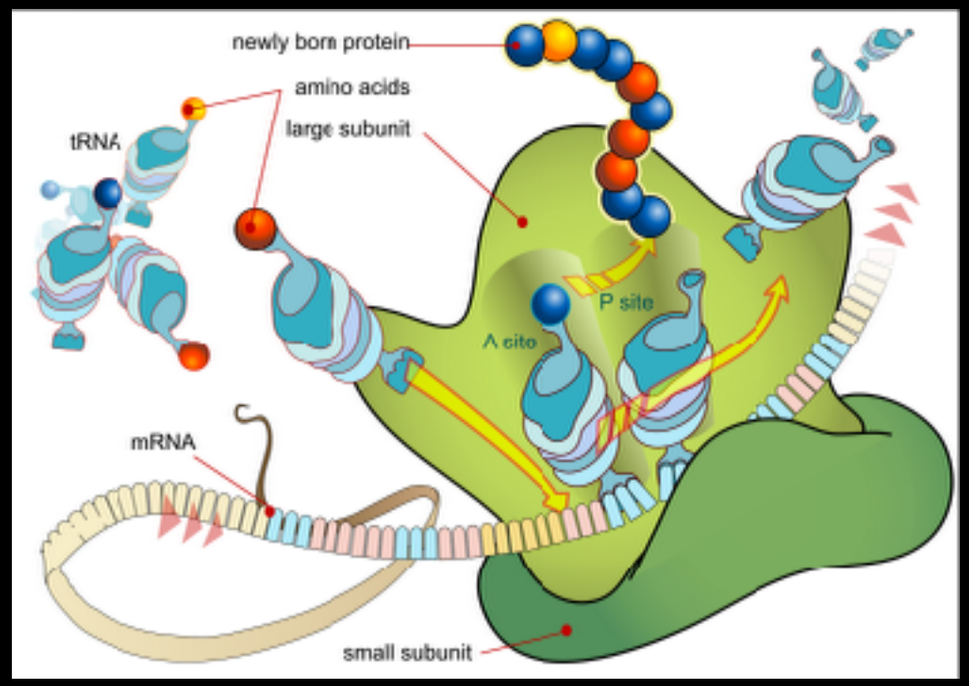
DNA

1

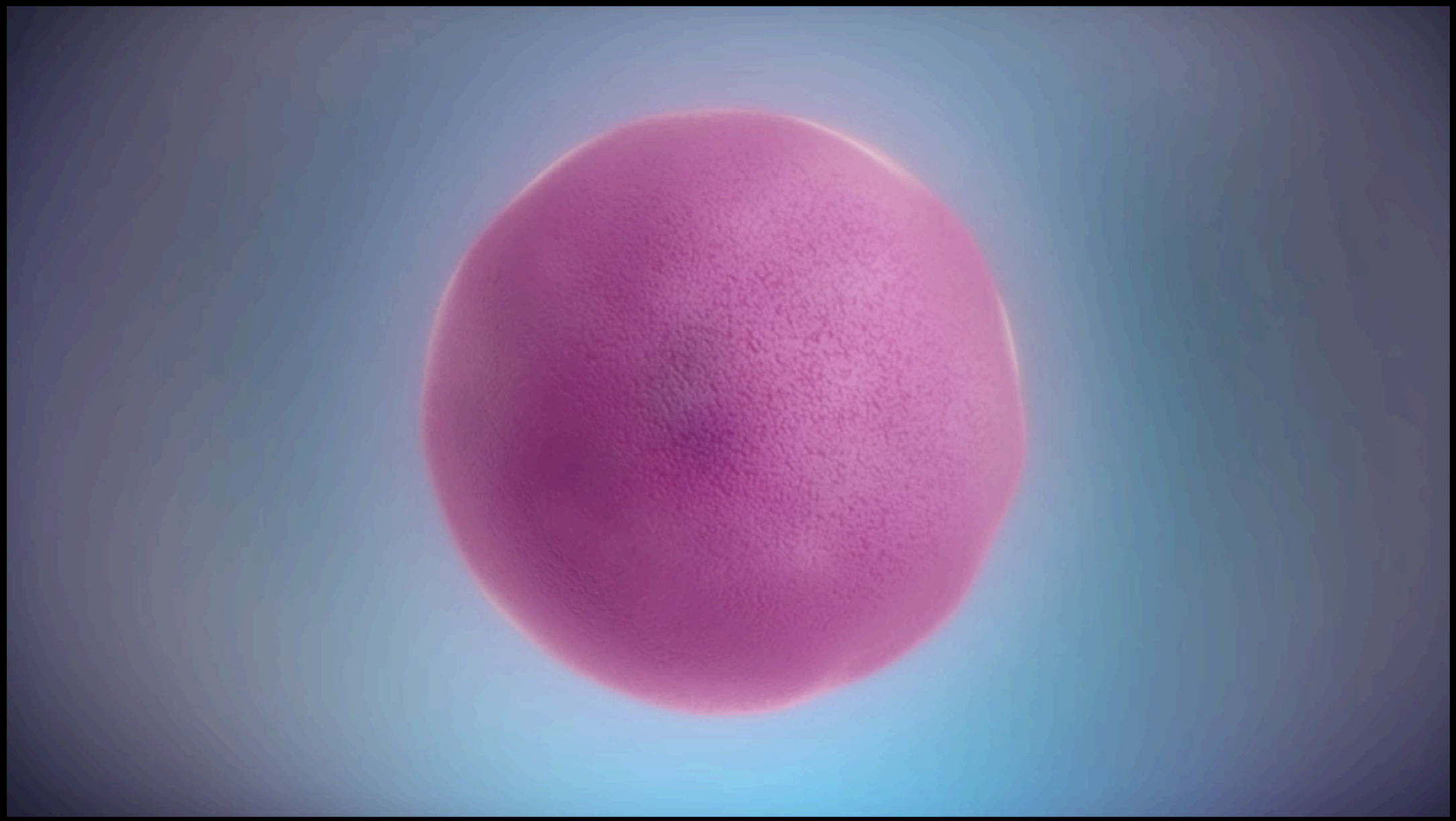
RNA

2

Protein



1. Transcription
2. Translation

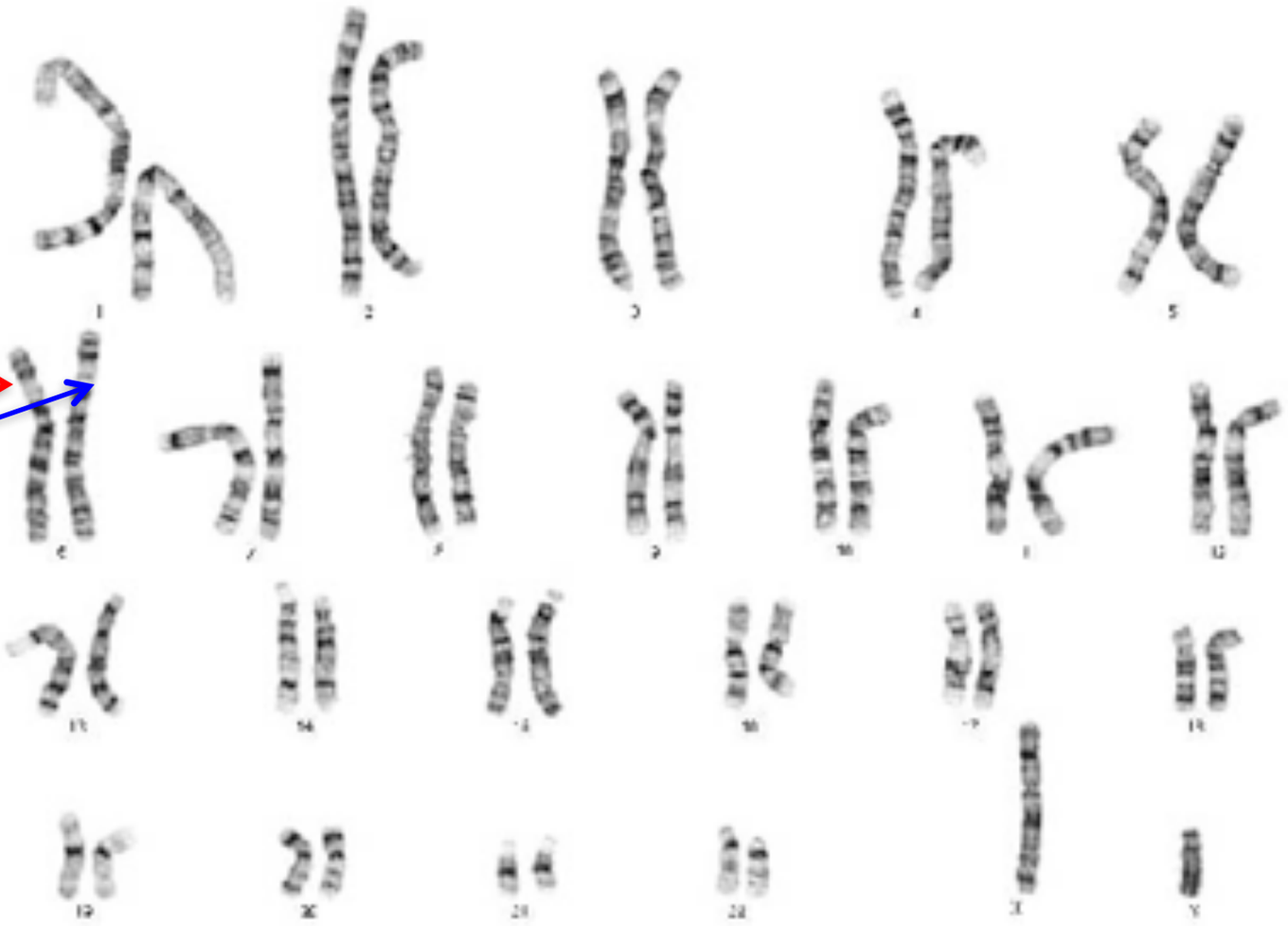


Structure of the Human Genome

Allele

acg**C**taga

acg**G**taga



22 pairs + X/Y

Structure of the Human Genome

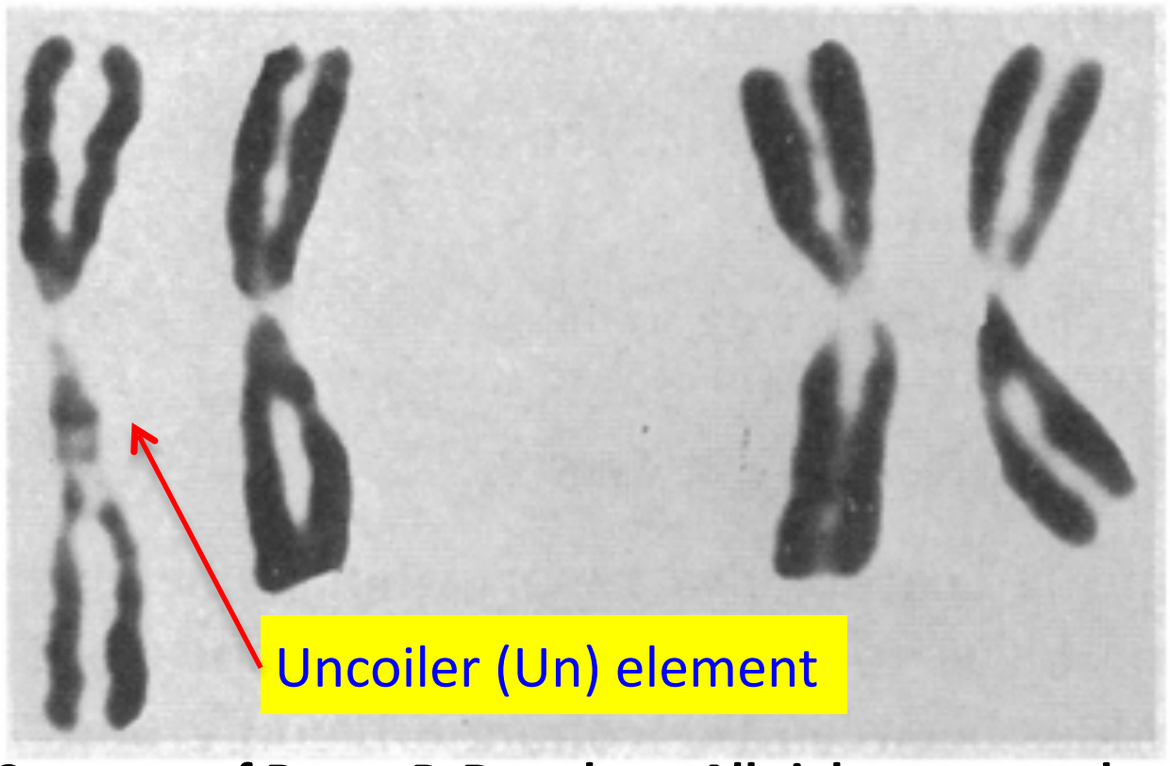
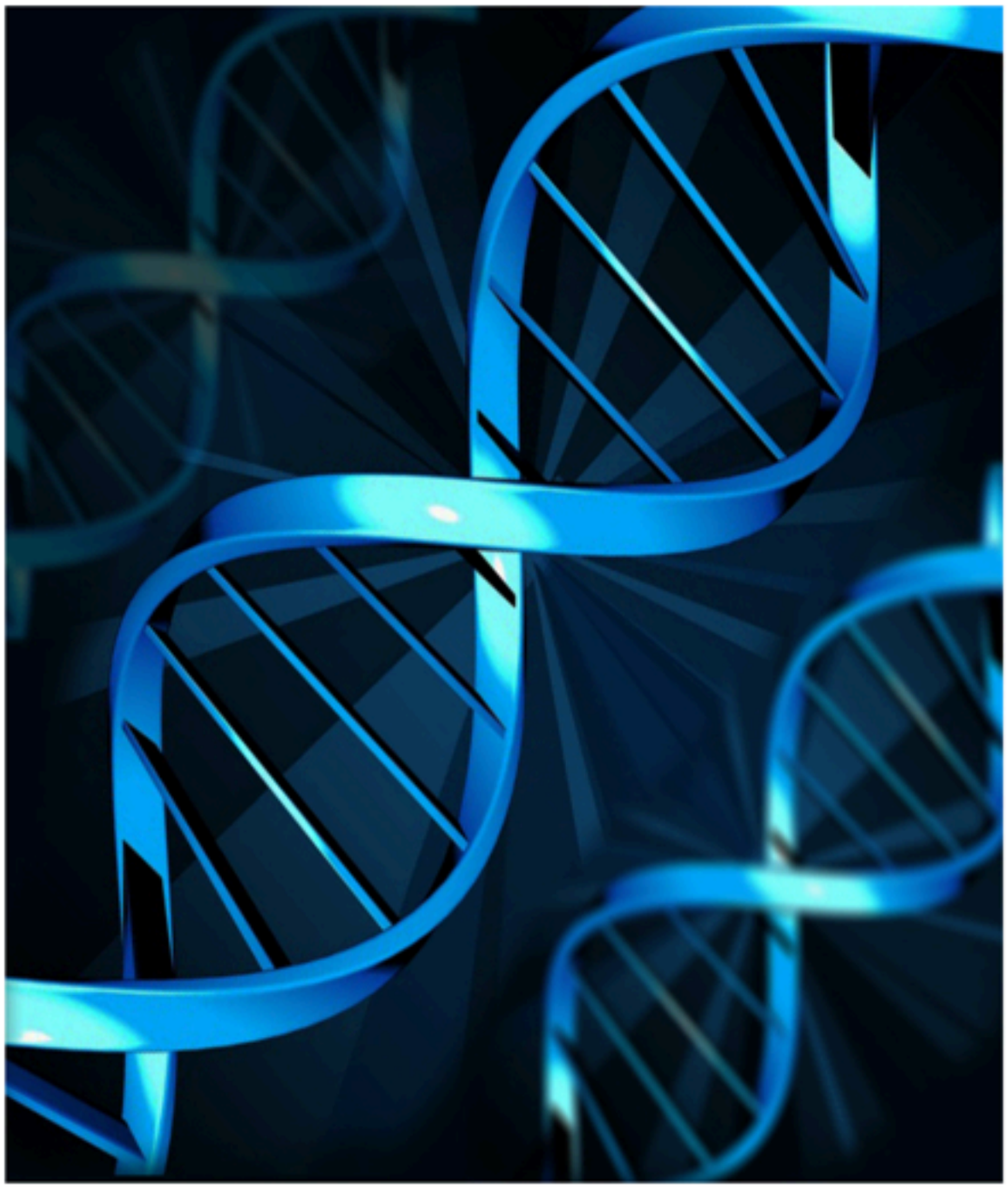


Figure 1: Two pairs of chromosome 1 from metaphase lymphocytes of family V11DA.
A pair of chromosomes from the proband's uncle is shown at left; the uncle is heterozygous for the variant chromosome, which is the leftmost member of the pair. A pair of chromosomes from the proband's sister is shown at right; the sister lacks the variant chromosome.

Chial, H. (2008) Gene mapping and disease. Nature Education 1(1):50

UNIVERSITY OF CALIFORNIA SANTA CRUZ UCSC **Genome Browser**

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us



Our tools

- **Genome Browser**
  interactively visualize genomic data
- **BLAT**
  rapidly align sequences to the genome
- **Table Browser**
  download data from the Genome Browser database
- **Variant Annotation Integrator**
  get functional effect predictions for variant calls
- **Data Integrator**
  combine data sources from the Genome Browser database
- **Gene Sorter**
  find genes that are similar by expression and other metrics
- **Genome Browser in a Box (GBiB)**
  run the Genome Browser on your laptop or server
- **In-Silico PCR**
  rapidly align PCR primer pairs to the genome
- **LiftOver**
  convert genome coordinates between assemblies
- **VisiGene**
  interactively view in situ images of mouse and frog

More tools...

http://genome-asia.ucsc.edu

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

http://genome-asia.ucsc.edu

Gene Structure

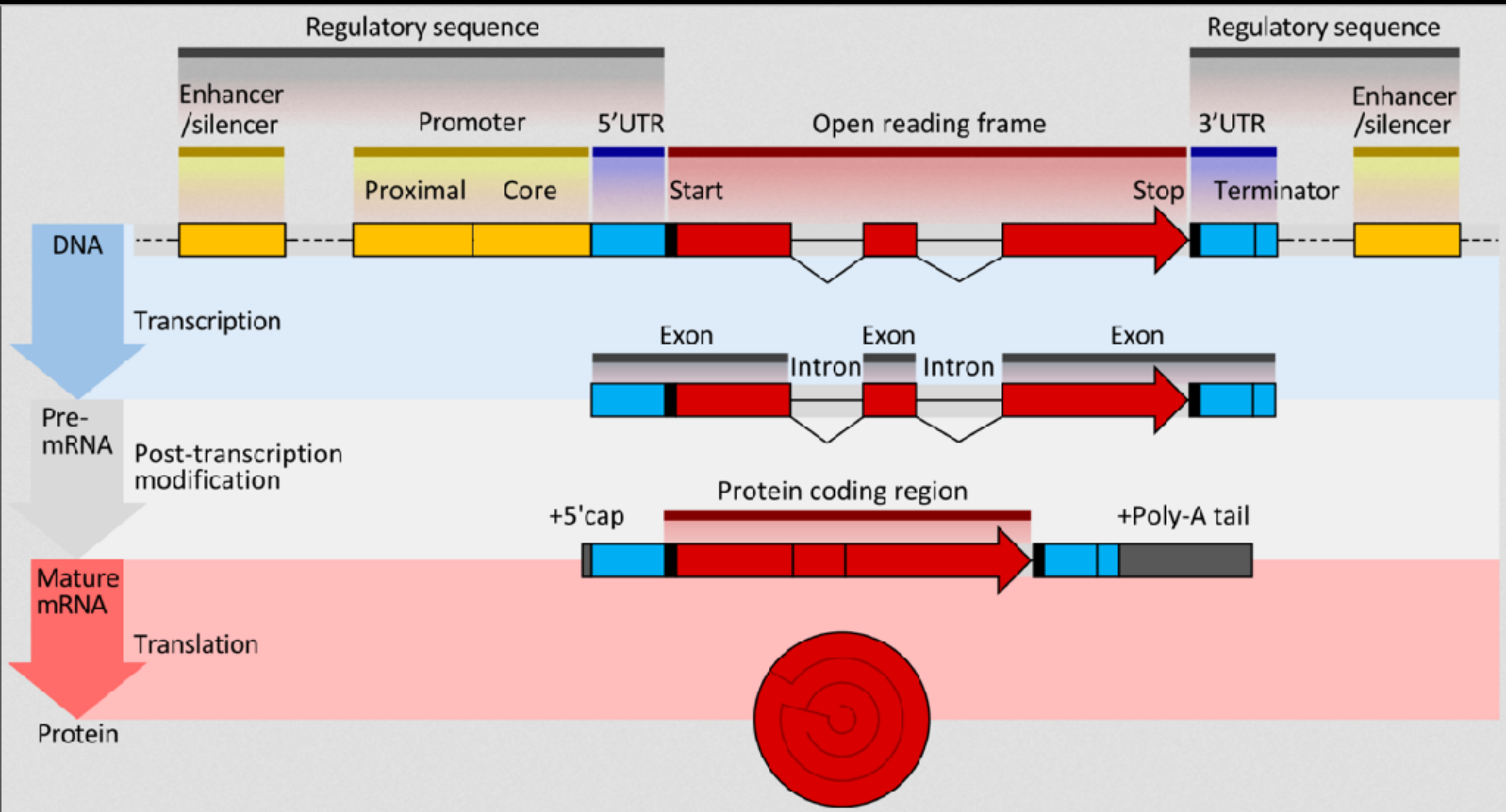


Image: Wikipedia

Regulation of Gene Expression

- Modification of DNA
  - Chromatin/Histone modification
- Regulation of transcription
  - CpG islands, transcription factors, repressor/activator, silencer/enhancer
- Post-transcriptional regulation (RNA life-cycle)
  - various nc-RNA, 3’UTRs
- Regulation of translation

Life science: many data types

Genes, genomes & variation

Gene, protein & metabolite expression

Protein sequences, families & motifs

Macromolecular structures

Chemogenomics & metabolomics

Interactions, reactions & pathways

Cross-domain tools & resources

Slide from Laura Emery (EMBL-EBI)

Bioinformatics Data

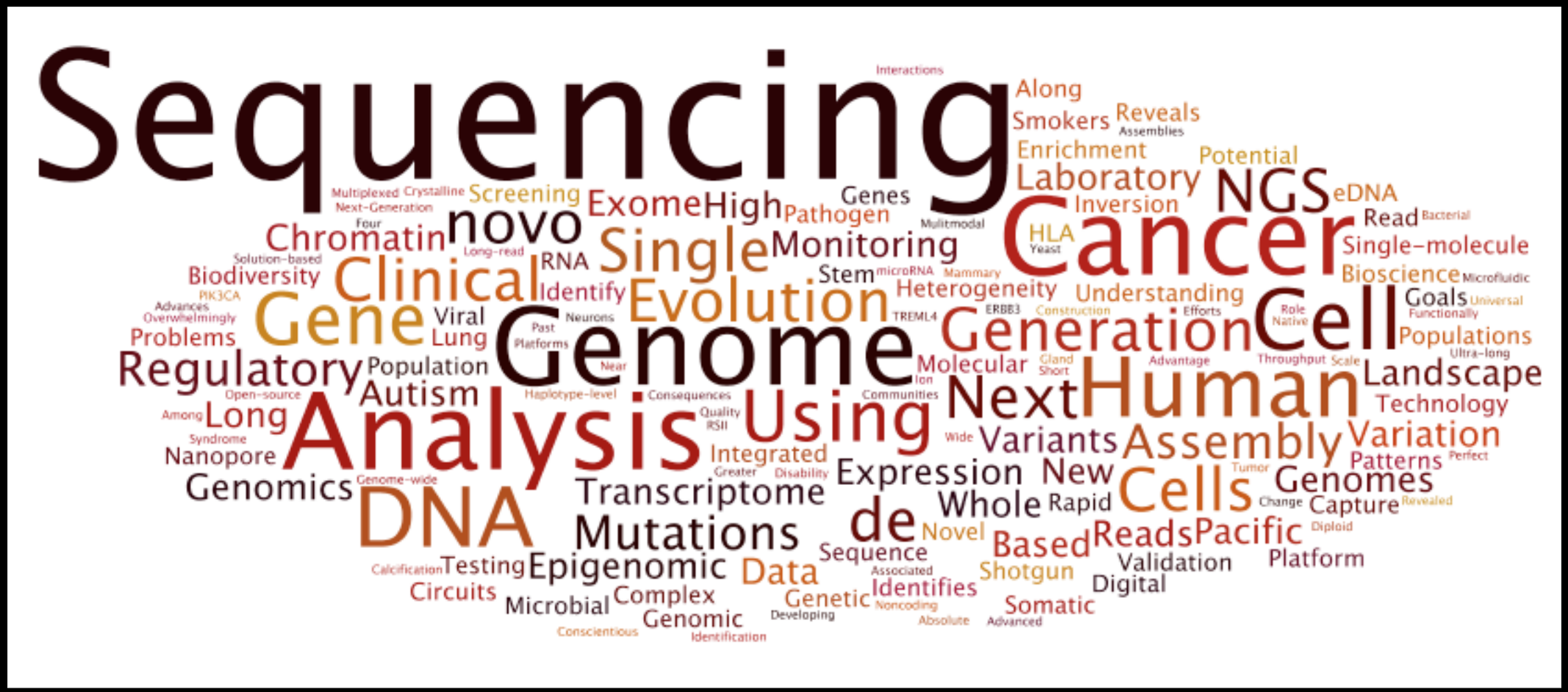


Image Credit: http://blog.illumina.com/blog/illumina/2014/02/10/marco-island-forecast

DNA Data

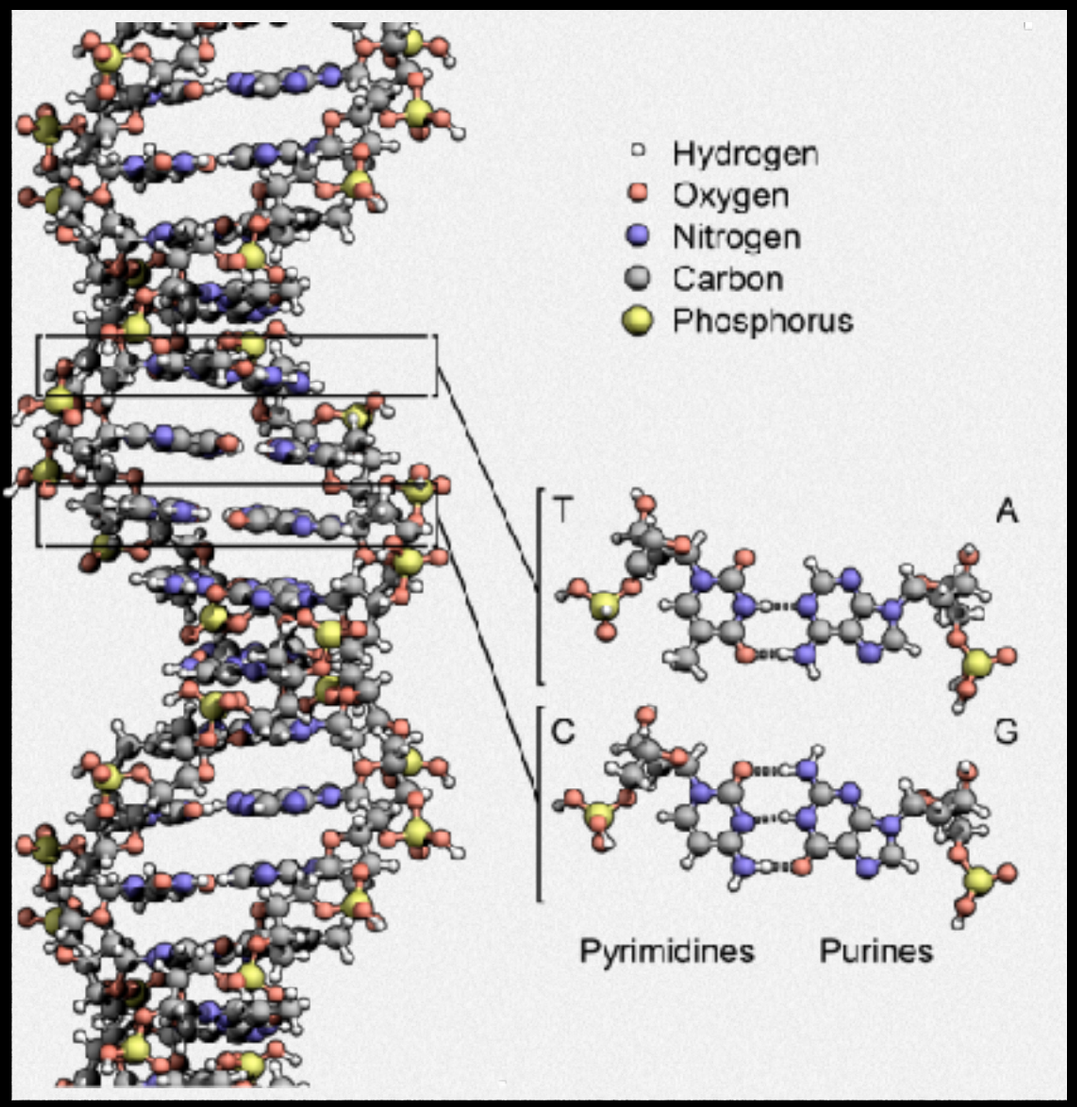


Image: Wikipedia

Representing Polymorphisms

Nucleic Acid Code	Meaning	Mnemonic
A	A	**A**denine
C	C	**C**ytosine
G	G	**G**uanine
T	T	**T**hymine
U	U	**U**racil
R	A or G	pu**R**ine
Y	C, T or U	p**Y**rimidines
K	G, T or U	bases which are **K**etones
M	A or C	bases with a**M**ino groups
S	C or G	**S**trong interaction
W	A, T or U	**W**eak interaction
B	not A (i.e. C, G, T or U)	**B** comes after A
D	not C (i.e. A, G, T or U)	**D** comes after C
H	not G (i.e., A, C, T or U)	**H** comes after G
V	neither T nor U (i.e. A, C or G)	**V** comes after U
N	A C G T U	**N**ucleic acid
X	masked	
-	gap of indeterminate length	

Sequence of Nucleotides

Homo sapiens chromosome 11, alternate assembly CHM1_1.1, whole genome shotgun sequence

NCBI Reference Sequence: NC_018922.2

GenBank Graphics

```
>gi|528476600:c5247236-5245631 Homo sapiens chromosome 11, alternate assembly
CHM1_1.1, whole genome shotgun sequence
ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCACCTGACTCCTGA
GGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGC
AGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGAGACAGAGAAG
ACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGG
TGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGG
CAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGAC
AACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACT
TCAGGGTGAGTCTATGGGACCCTTGATGTTTTCTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAG
GAAGGGGAGAAGTAACAGGGTACAGTTTAGAATGGGAAACAGACGAATGATTGCATCAGTGTGGAAGTCT
CAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTTGTTTAATTCTTGCTTTCT
TTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAAAGGAAATA
TCTCTGAGATACATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTGGAAT
ATATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTATTTTCTTTTATTTTTAATTGATACATAAT
CATTATACATATTTATGGGTTAAAGTGTAATGTTTTAATATGTGTACACATATTGACCAAATCAGGGTAA
TTTTGCATTTGTAATTTTAAAAAATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATA
CTTTCCCTAATCTCTTTCTTTCAGGGCAATAATGATACAATGTATCATGCCTCTTTGCACCATTCTAAAG
AATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATTTCTGCATATAAATATTTCTGCATATAAAT
TGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGCTACAATCCAGCTACCATTCTGCTTTTATTTT
ATGGTTGGGATAAGGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCATACCTCTT
ATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCA
CCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCA
CTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACT
GGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGC
```

FASTA (Nucleotide Sequence)

Sequence Data

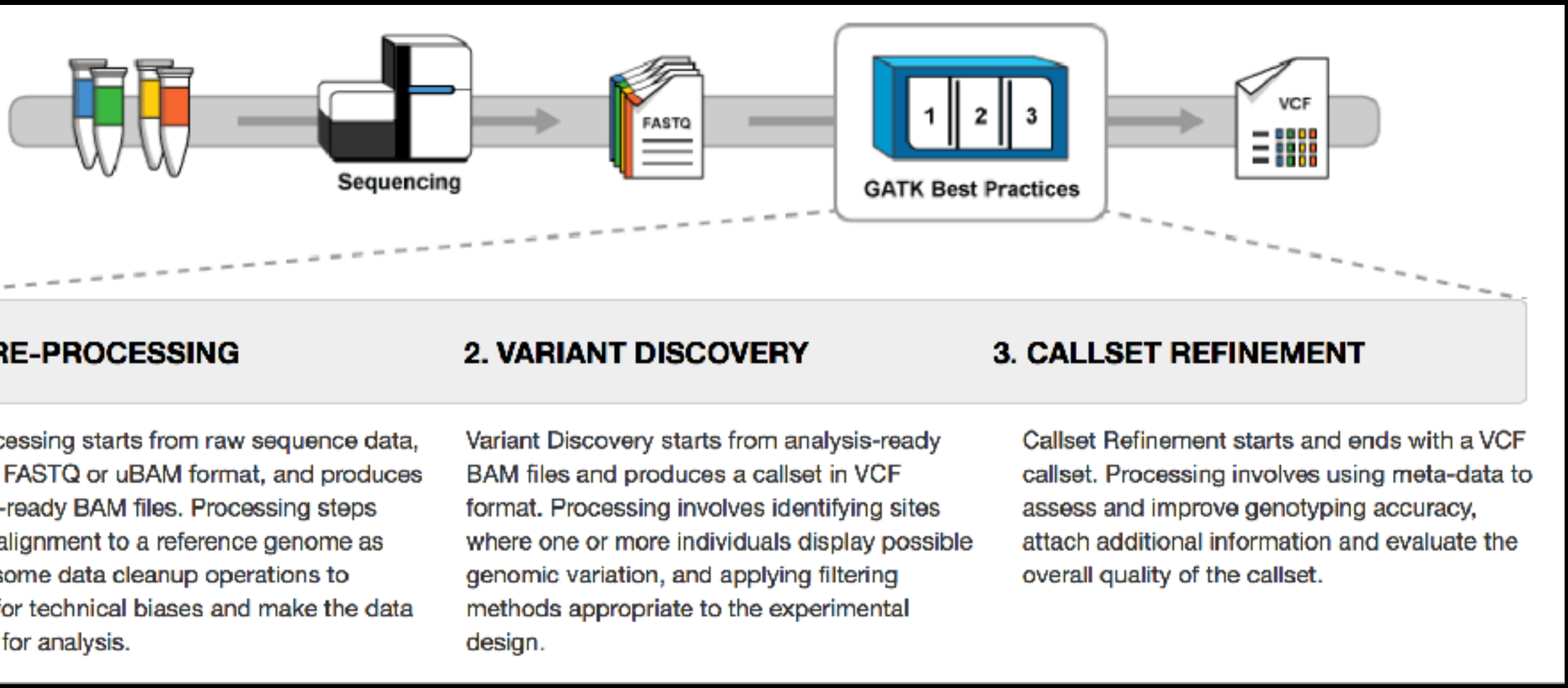
```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

FASTA (Protein Sequence)

Extension	Meaning	Notes
fasta (.fas)	generic fasta	Any generic fasta file. Other extensions can be fa, seq, fsa
fna	fasta nucleic acid	Used generically to specify nucleic acids.
ffn	FASTA nucleotide of gene regions	Contains coding regions for a genome.
faa	fasta amino acid	Contains amino acids. A multiple protein fasta file can have the more specific extension mpfa.
frn	FASTA non-coding RNA	Contains non-coding RNA regions for a genome, in DNA alphabet e.g. tRNA, rRNA

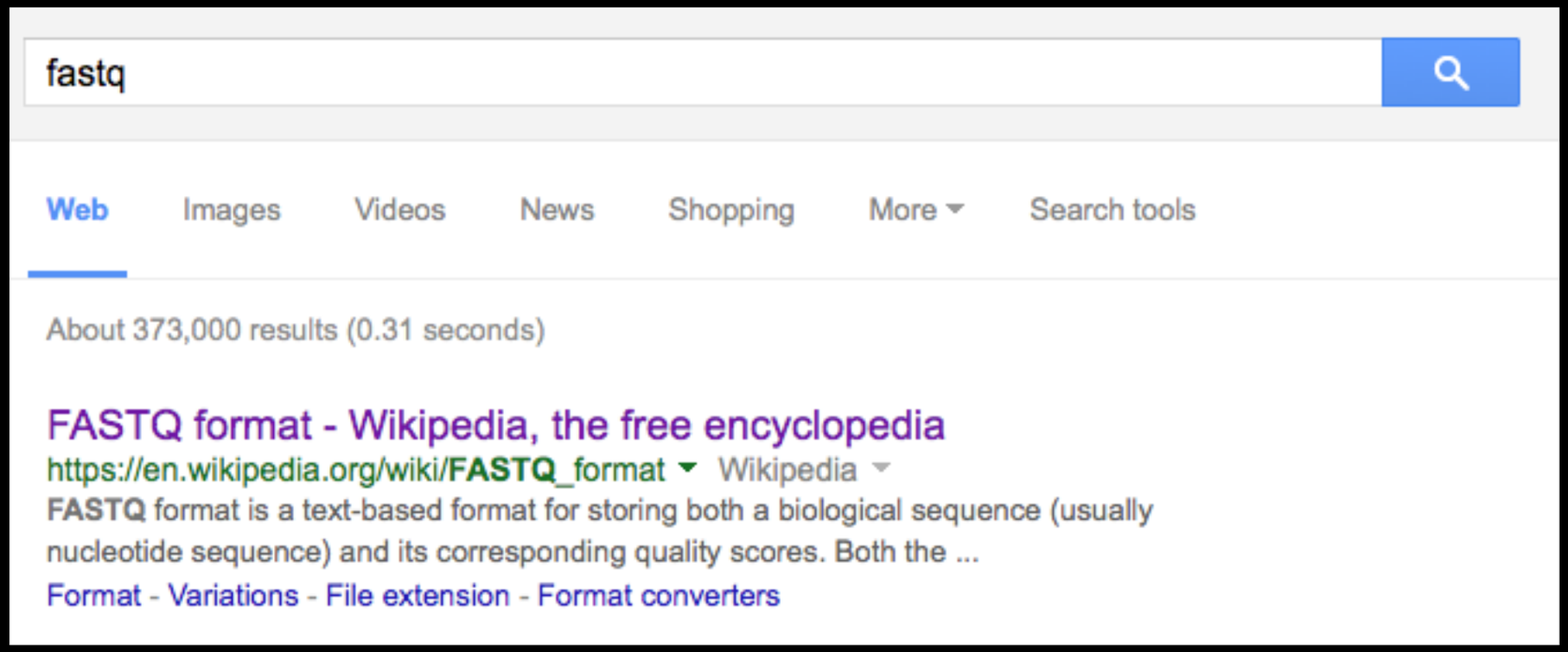
FASTQ ?!

Next-Generation Sequencing



https://www.broadinstitute.org/gatk/guide/best-practices

What's that *FASTQ* format?



A FASTQ file normally uses four lines per sequence.

Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).

Line 2 is the raw sequence letters.

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

What's that *FASTQ* format?

A FASTQ file normally uses four lines per sequence.

- **Line 1** begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).
- **Line 2** is the raw sequence letters.
- **Line 3** begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.
- **Line 4** encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

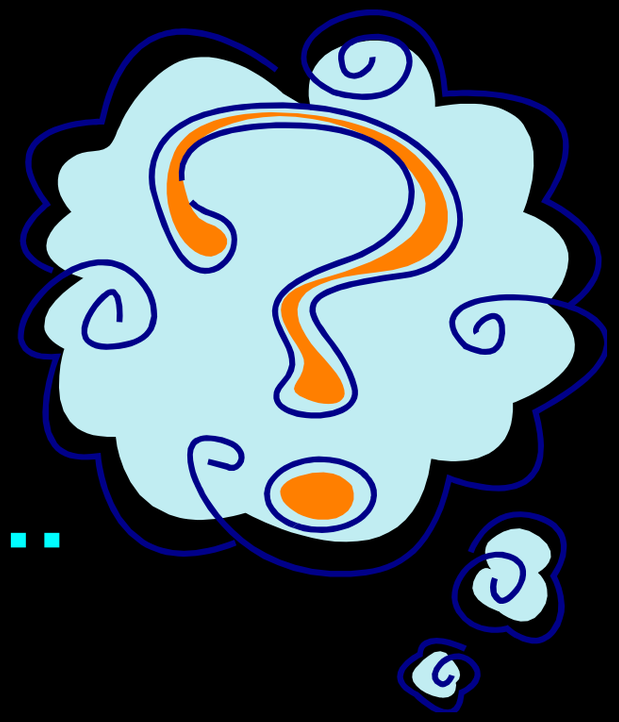
Quality Score

Tips for Handling of Biological Data

1. Understand how data were generated (sources of errors)
2. Check for the errors (quality control)
3. Clarify the research questions (hints on the required tools)
4. Read the manuals! (all the how-to stuff)
5. Understand the context of the research (interpretation of the results)

6. **Stay up-to-date!!**

Examples



genetic epidemiology…

1) Genome-wide Association Study using SNP array

2) Finding Disease Genes with Exome Sequencing Study

Genetic Epidemiology

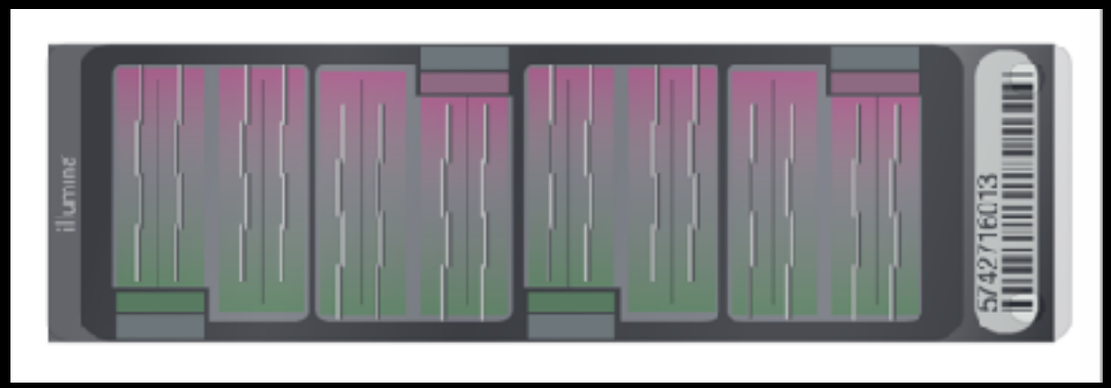
- **A** study of the role of genetic factors in determining health and disease in families and in population, and the **interplay** of such genetic factors with environmental factors.
- “A science which deals with the etiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in populations“ (Newton Morton, 1978)

SNP Array Data

Flagship array scanner

Optimized for high density HumanOmni BeadChips

- Highest sample throughput
- Industry leading data quality
- Flexible platform



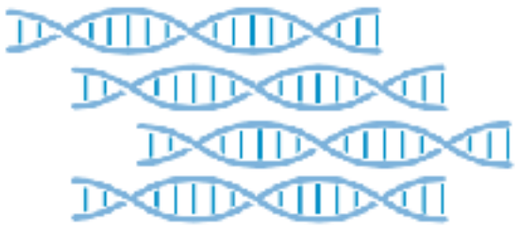
Illumina SNP Array

Affymetrics SNP Array

FIGURE 1: INFINIUM II ASSAY PROTOCOL

GENOMIC DNA (750 ng)

DAY 1



1 Make amplified DNA

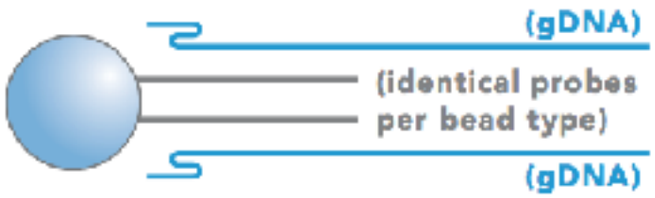
2 Incubate amplified DNA

DAY 2

3 Fragment amplified DNA

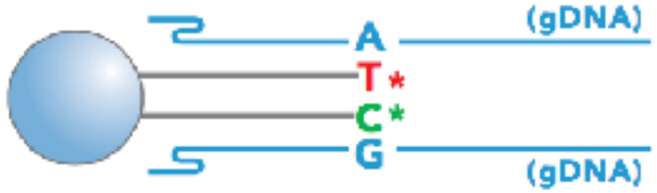
4 Precipitate & resuspend

5 Prepare BeadChip



6 Hybridize samples on BeadChip

DAY 3



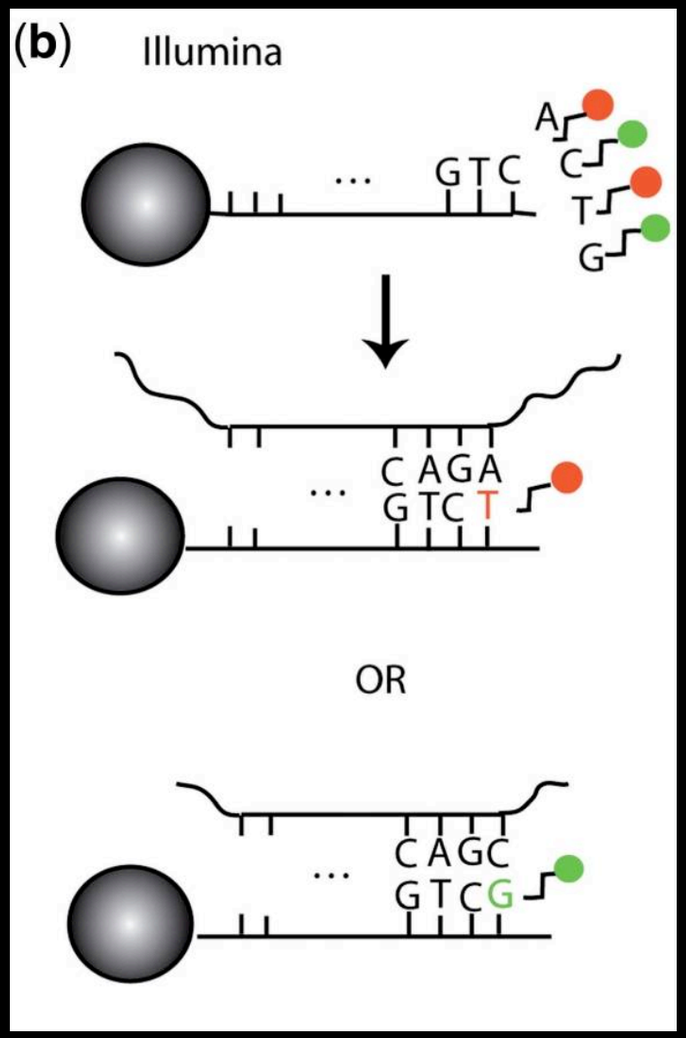
7 Extend/Stain samples on BeadChip

8 Image BeadChip

9 Auto-call genotypes and generate reports

*Indicates stain in red channel
*Indicates stain in green channel

Illumina® SNP Array



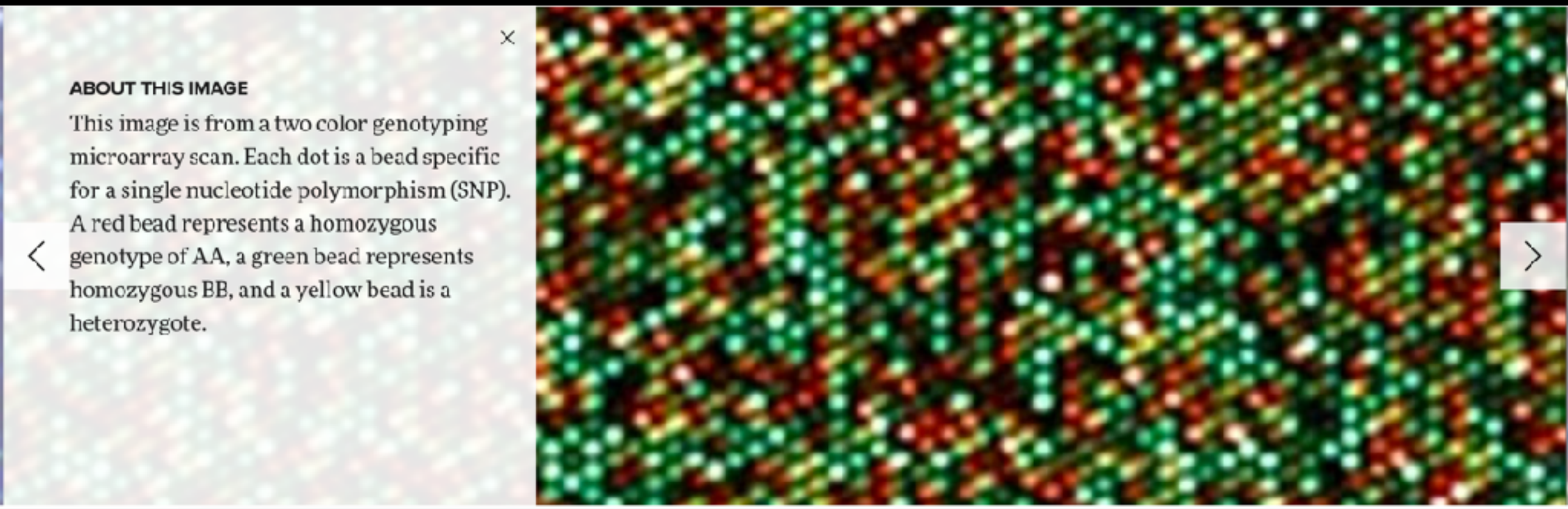
50-mer complimentary sequence

Specific *single-base* extension

Specific *single-base* extension

Nucleic Acids Research, 2009, Vol. 37, No. 13

Raw Beadchip Image

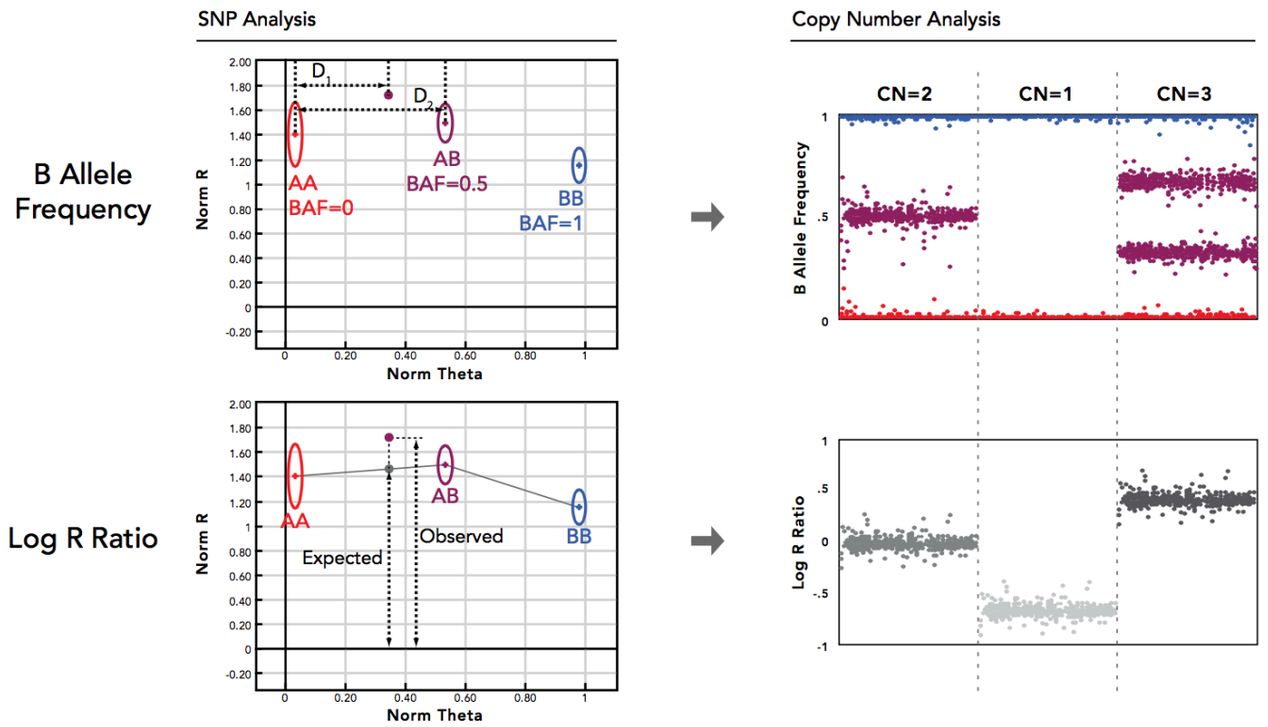


Data Conversion

- Image Processing: converting image to intensity data
- Genotype Calling: clustering intensity data to genotype data or Copy number variation

Intensity Data
B-allele Frequency & Log-R Ratio

Figure 3: Calculation of Log R Ratio and B Allele frequency

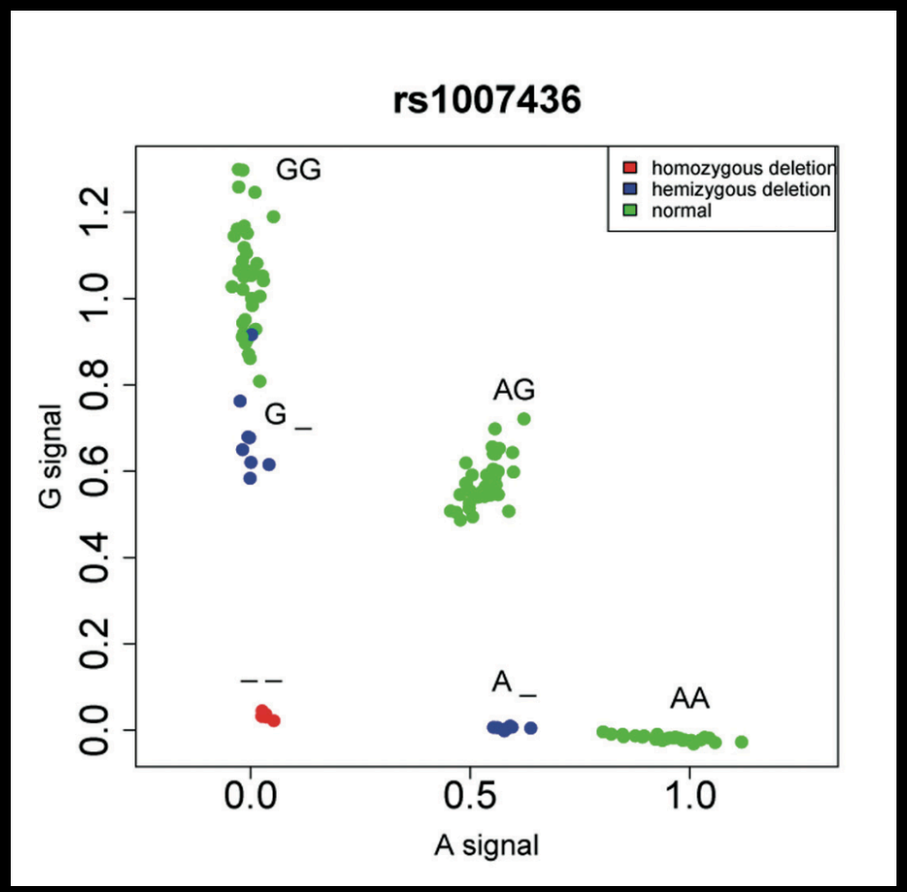


The allelic copy ratio in terms of B allele frequency (BAF) is calculated from the θ value of a sample and the expected cluster positions (ellipses, left top panel). The allele frequency is determined as a linear interpolation in the θ-dimension related to the allele frequency of each cluster (0.0, 0.5, and 1.0). In this example, a data point (purple dot) falling approximately 2/3 of the distance from the AA to the AB cluster (D1/D2 = .66) has an allele frequency of 0.33 (0.66 * 0.5).
The $\log_2$ R ratio is calculated as the ratio between observed normalized intensity of the experimental sample to the expected intensity (left bottom panel). The expected intensity is determined as a linear interpolation as a function of the sample θ (grey line) of the expected cluster positions (ellipses).
These two transformed parameters, B allele frequency and $\log_2$ R ratio, are then plotted along the entire genome for all SNPs on the array (right panel). The plot of these two parameters exhibits diagnostic signature profiles of copy number (example copy numbers 2, 1, and 3 shown) and specific classes of structural variation.
Figure adapted from Peiffer, et. al, 2006.

Illumina Technical Note: Interpreting Infinium Assay Data for Whole-Genome Structural Variation

Genotype Calling

- Cluster intensity data to genotype cluster



Normalized
Allelic intensity ratios (θ)
Intensity values (R)

Genotype = a genetic type of an organism

Allele

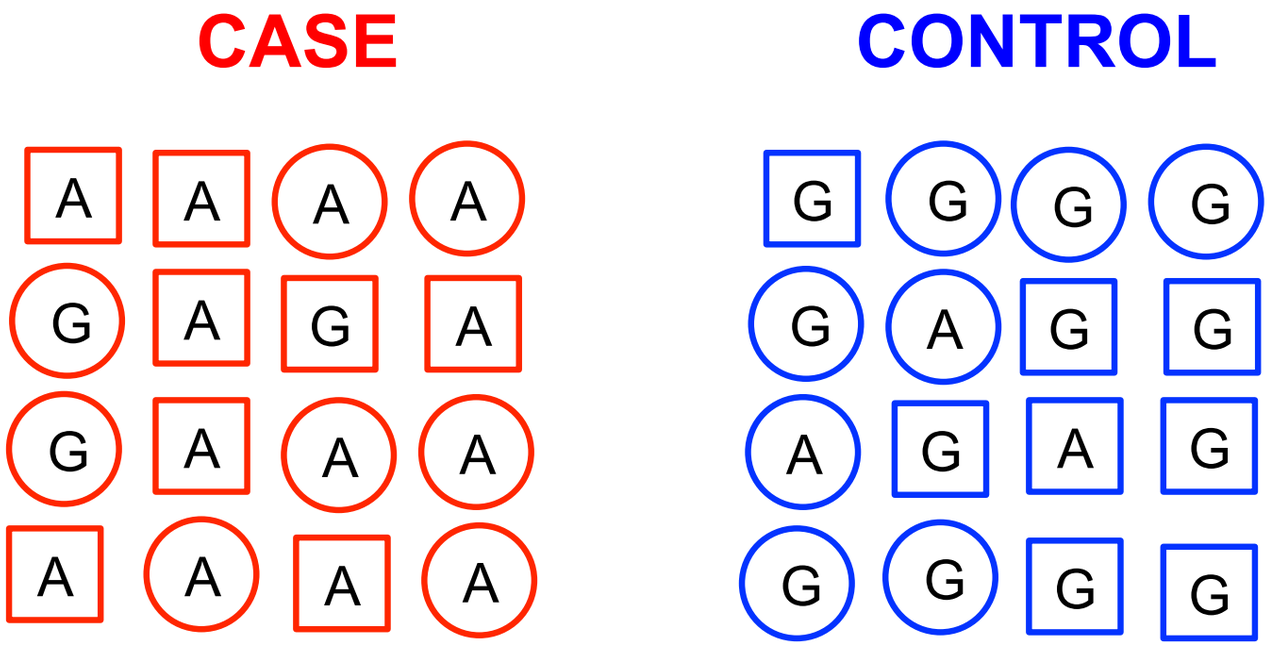
Haplotype = One copy of a genotype

Diplotype = Two copies of a genotype

Nucleic Acids Research, 2009, Vol. 37, No. 13

BEADCHIP	ARRAY FORMAT/ SAMPLE THROUGHPUT	MARKERS PER SAMPLE	CUSTOM MARKER ADD-ON CAPABILITY	DATA SET USED FOR CONTENT SELECTION	CONTENT DESCRIPTION
HumanOmni5-Quad	4/460	4.3 million	500,000	1000 Genomes Dec 2012 release (MAF ≥ 1%)	Access the highest value content, plus 500K of your own.
HumanOmni5Exome	4/460	~ 4.5 million	200,000	1000 Genomes Dec 2012 release (MAF ≥ 1%), plus exome content selected from 12,000 individual exome sequences taken from various large sequence projects	The most powerful microarray. 4.3 million whole-genome variants down to 1% MAF, combined with novel functional exonic variants taken from over 12,000 sequenced exomes.
HumanOmni2.5-8	8/1067	~2.5 million	200,000	1000 Genomes Project Pilot (MAF ≥ 2.5%)	Common and rare variants targeting down to 2.5% MAF selected from the 1000 Genomes Project
HumanOmni2.5Exome-8	8/1067	~ 2.6 million	N/A	1000 Genomes Project Pilot (MAF > 2.5%), plus exome content selected from 12,000 individual exome sequences taken from various large sequence projects	Ccomprehensive coverage and functional exonic content for next generation genotyping, GWAS, and CNV analysis.
HumanOmniExpress-24	24/2800	> 715,000	50,000	HapMap MAF ≥ 5%	Scan thousands of samples per week using optimized common tag SNPs.

Genetic Association Study



"A" allele is associated with *CASE status*

Genotype = a genetic type of an organism

Allele

Haplotype = One copy of a genotype

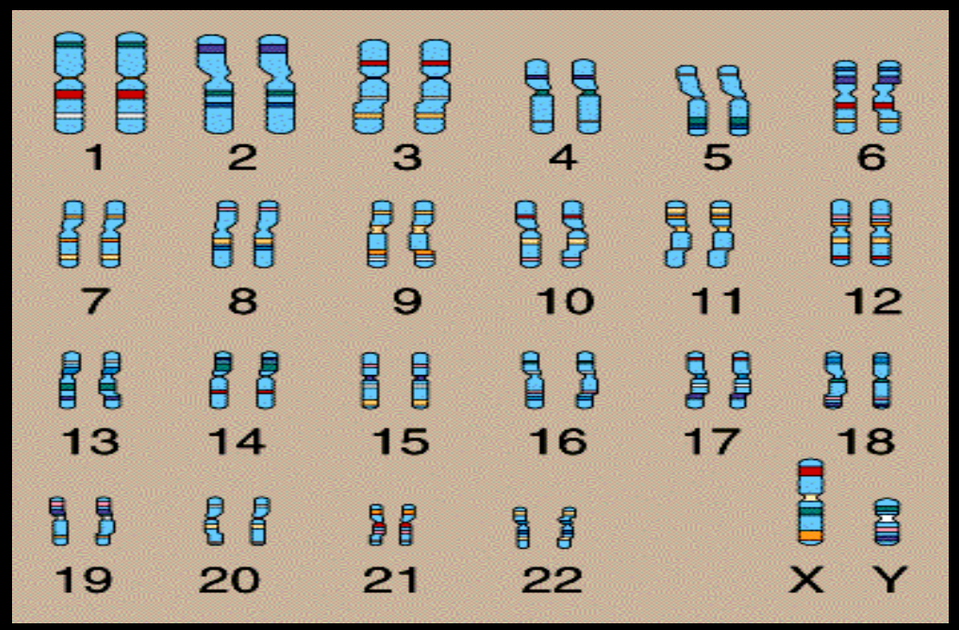
Diplotype = Two copies of a genotype

Locating causative genes in the genome

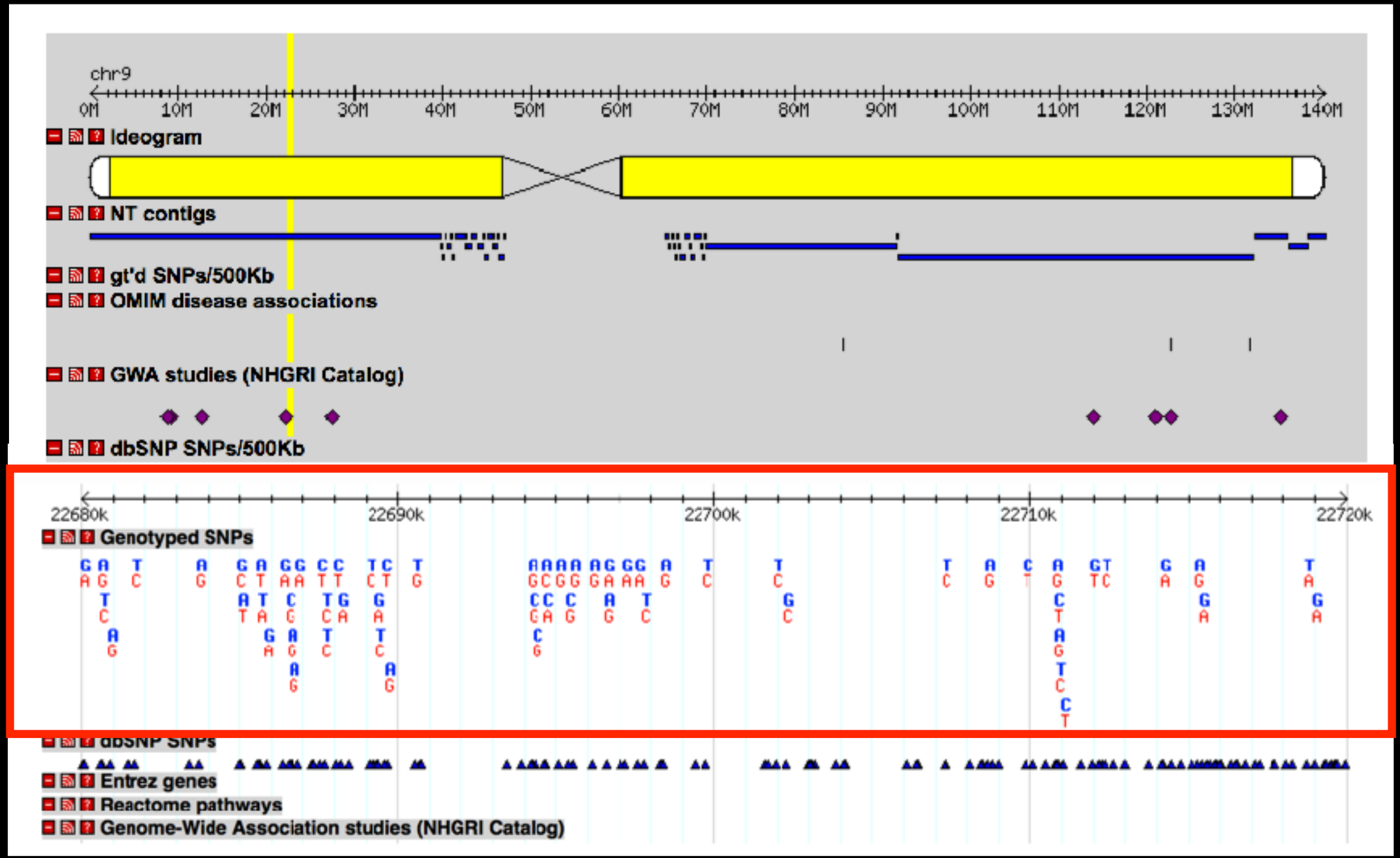
Gene = A functional unit in a genome

- 24 chromosomes
- 3 billions base-pair genome
- ~ 20,000 genes

- How can we find these genes?



SNP Array Data



http://hapmap.ncbi.nlm.nih.gov/cgi-perl/gbrowse/hapmap28_B36/#search

A Common Allele on Chromosome 9 Associated with Coronary Heart Disease

Ruth McPherson,[1]*† Alexander Pertsemlidis,[2]* Nihan Kavaslar,[1] Alexandre Stewart,[1] Robert Roberts,[1] David R. Cox,[3] David A. Hinds,[3] Len A. Pennacchio,[4,5] Anne Tybjaerg-Hansen,[6] Aaron R. Folsom,[7] Eric Boerwinkle,[8] Helen H. Hobbs,[2,9] Jonathan C. Cohen[2,10]†

Coronary heart disease (CHD) is a major cause of death in Western countries. We used genome-wide association scanning to identify a 58-kilobase interval on chromosome 9p21 that was consistently associated with CHD in six independent samples (more than 23,000 participants) from four Caucasian populations. This interval, which is located near the *CDKN2A* and *CDKN2B* genes, contains no annotated genes and is not associated with established CHD risk factors such as plasma lipoproteins, hypertension, or diabetes. Homozygotes for the risk allele make up 20 to 25% of Caucasians and have a ~30 to 40% increased risk of CHD.

McPherson, et al. Science 316:1488-1491 (2007).

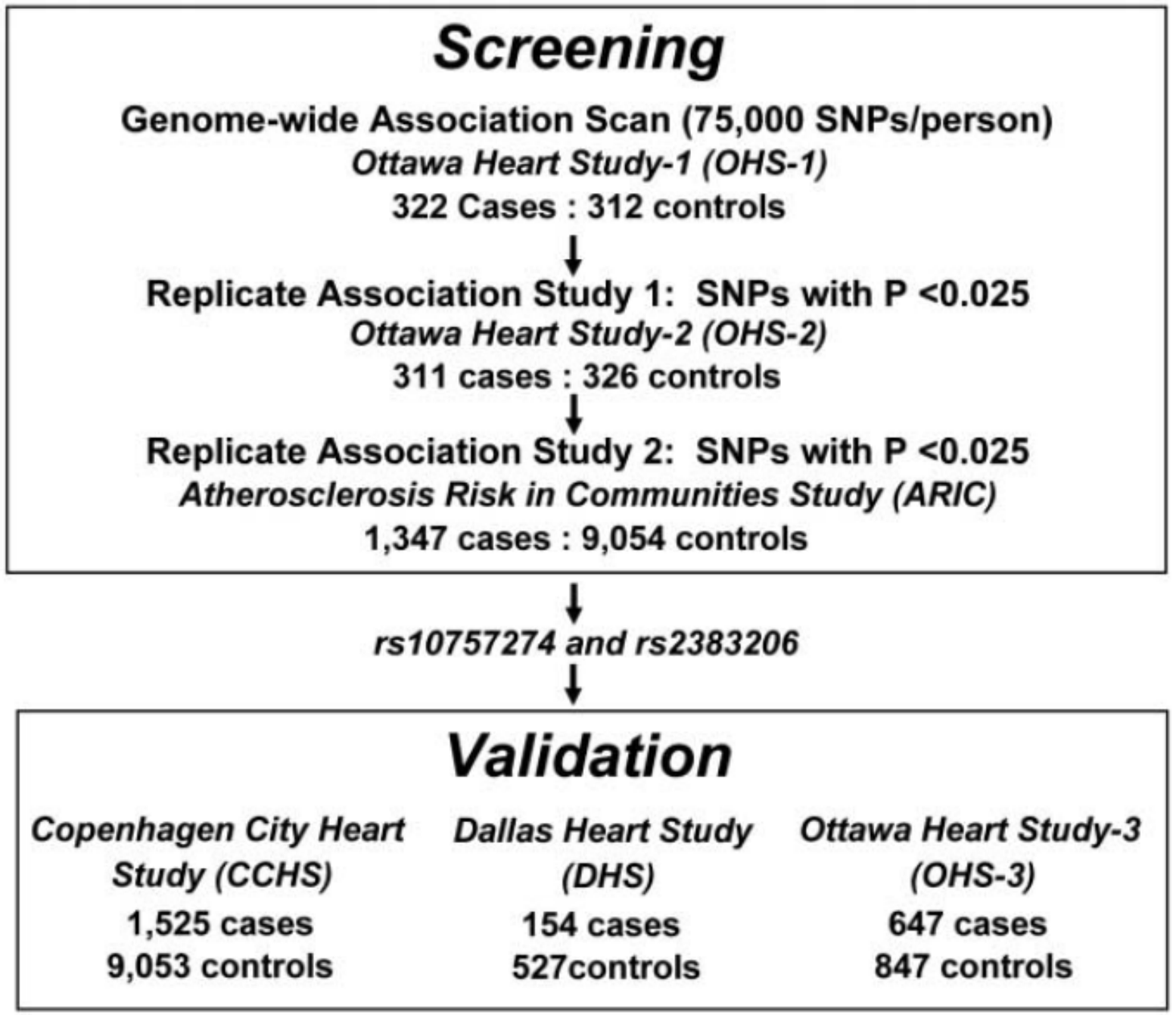
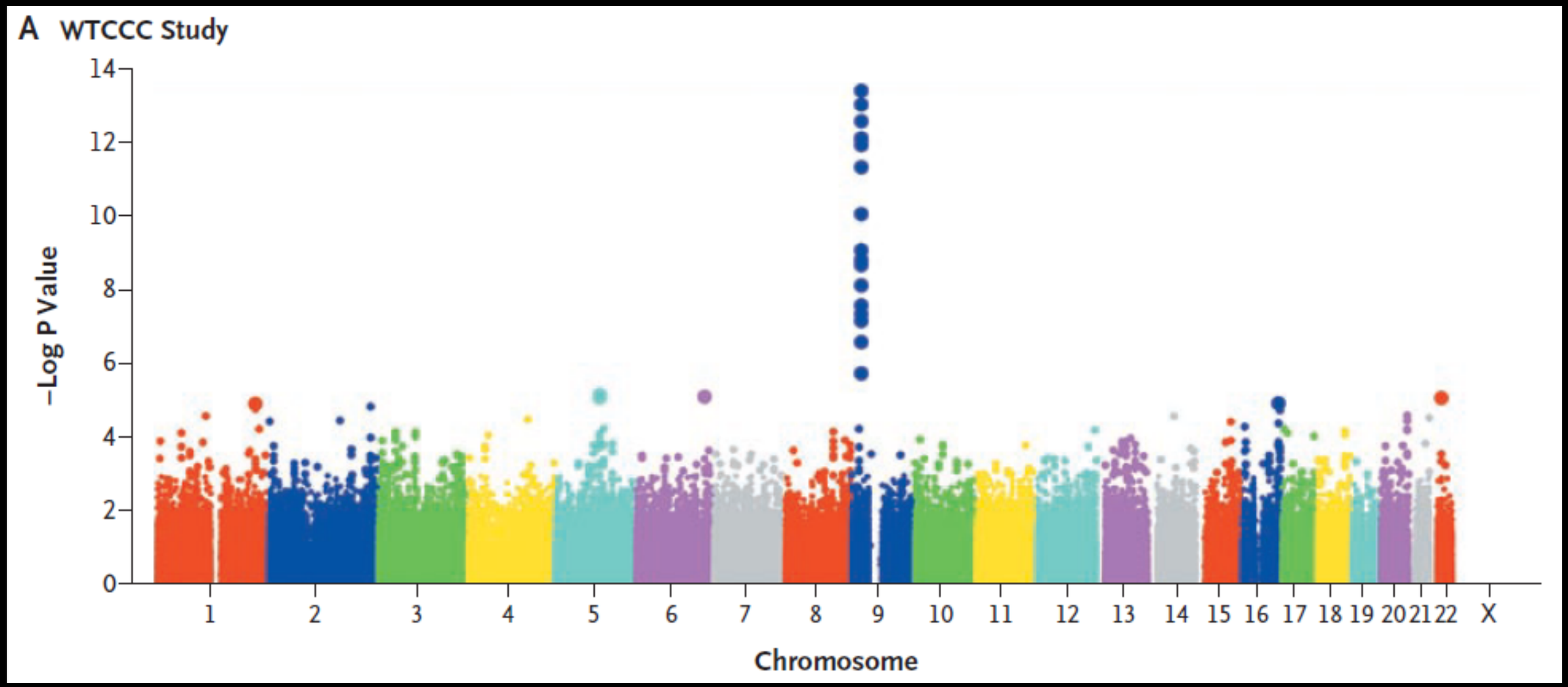


Fig. 1. Study design for identification and validation of sequence variants associated with CHD. Assuming independence, the probability of any single SNP achieving a nominal significance level of 0.025 in all three studies with the associations being in the same direction was 3.9×10^{-6} ($0.025^3 \times 0.5^2$); thus, none of the 100,000 SNPs would be expected by chance to replicate consistently in all three comparisons.

Summary of the results



Samani, et al. N Engl J Med 357(5): 443-53 (2007).

Gain-of-function mutations in the phosphatidylserine synthase 1 (*PTDSS1*) gene cause Lenz-Majewski syndrome

Sérgio B Sousa[1,2], Dagan Jenkins[3,18], Estelle Chanudet[4,18], Guergana Tasseva[5,18], Miho Ishida[1], Glenn Anderson[6], James Docker[7], Mina Ryten[8,9], Joaquim Sa[2], Jorge M Saraiva[2,10], Angela Barnicoat[11], Richard Scott[11], Alistair Calder[12], Duangrurdee Wattanasirichaigoon[13], Krystyna Chrzanowska[14], Martina Simandlová[15], Lionel Van Maldergem[16,17], Philip Stanier[7], Philip L Beales[3,4], Jean E Vance[5] & Gudrun E Moore[1]

Lenz-Majewski syndrome (LMS)

History

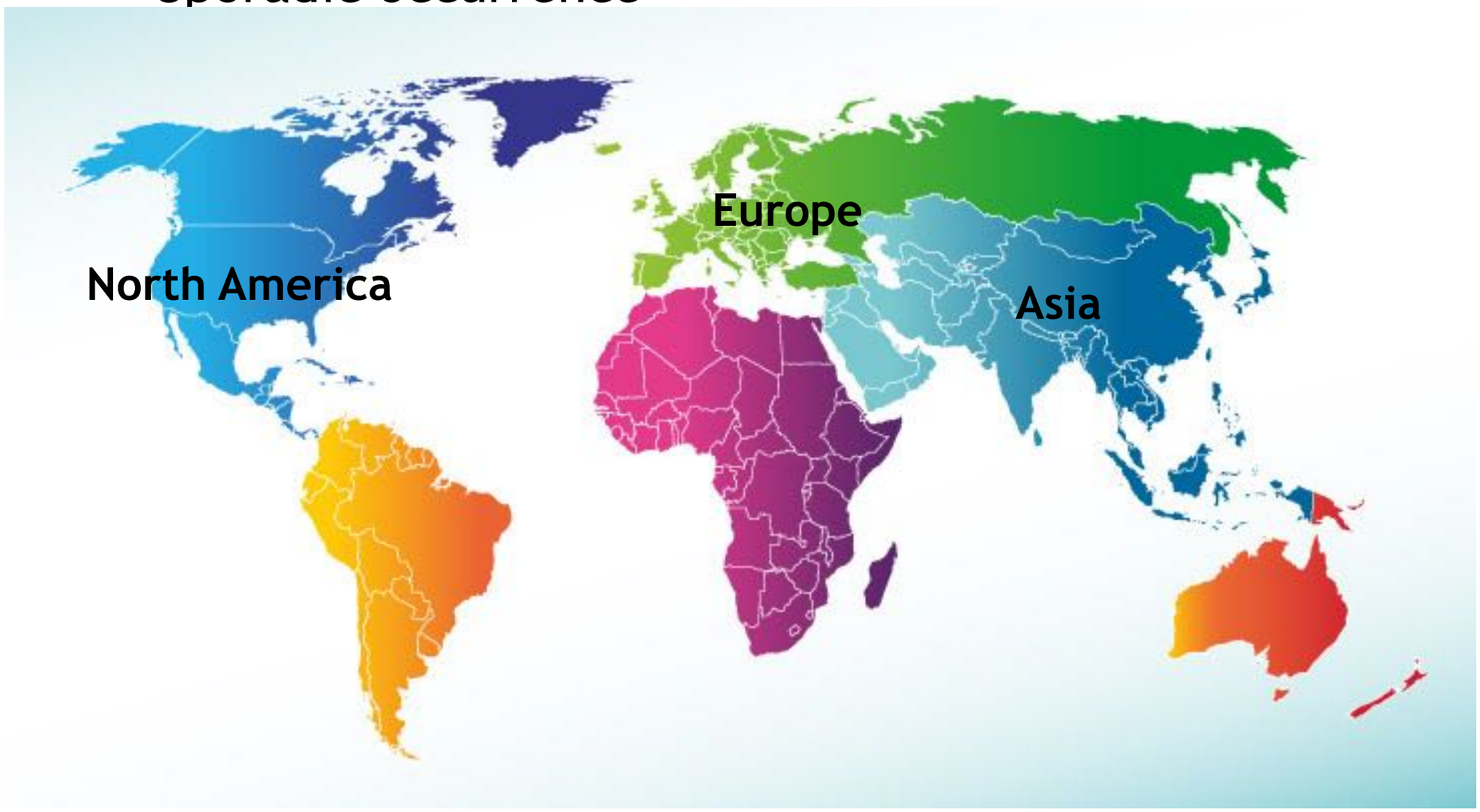
- First described by Braham in 1969
    - thought Camurati-Engelmann's syndrome
- Lenz and Majewski in 1974
    - suggested new syndrome
- Robinow in 1977
    - named **Lenz-Majewski hyperostotic dwafism**
    - " a syndrome of multiple congenital anomalies, mental retardation, and progressive skeletal sclerosis"

(Majewski, 2000)

Lenz-Majewski syndrome (LMS)

Incidence

- A very rare disorder : 10 cases have been reported
    - 6 males, 4 females
- Sporadic occurrence



Lenz-Majewski syndrome (LMS)

Clinical features

- Intellectual disability
- Craniofacial anomalies

  - large ears

(Gorlin and Whitley, 1983)

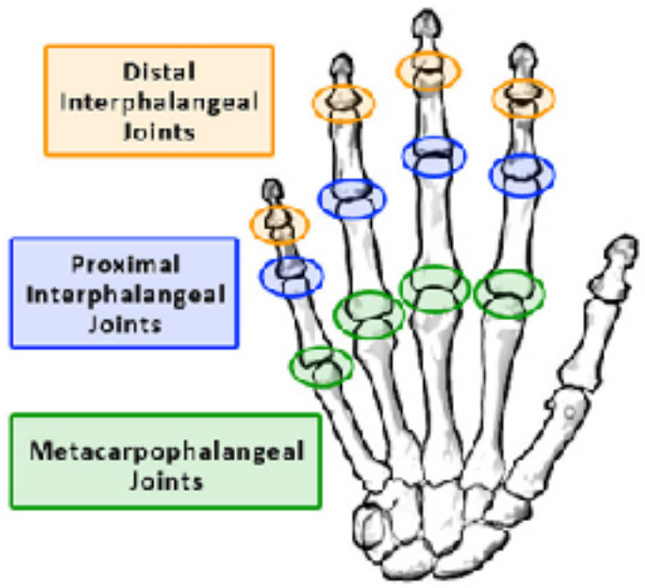
(Majewski, 2000)

  - broad/prominent forehead
  - hypertelorism
    (increased distance between the eyes)

Lenz-Majewski syndrome (LMS)

Clinical features

- loose and wrinkled atrophic skin/ cutis laxa



(Shoja *et al*. 2012)

- Distal limbs anomalies

  - Brachydactyly (shortness of the fingers and toes)

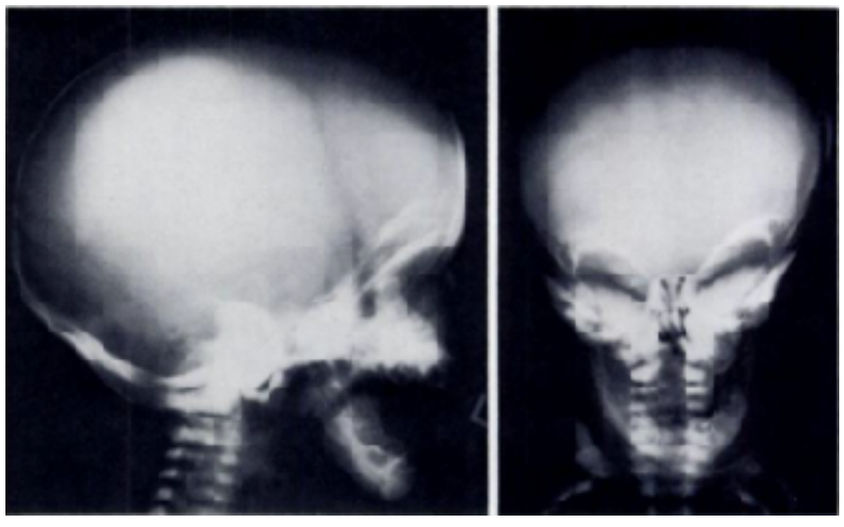
  - Cutaneous syndactyly (fusion of fingers)

  - Proximal symphalangism (fusion of the proximal interphalangeal joints)

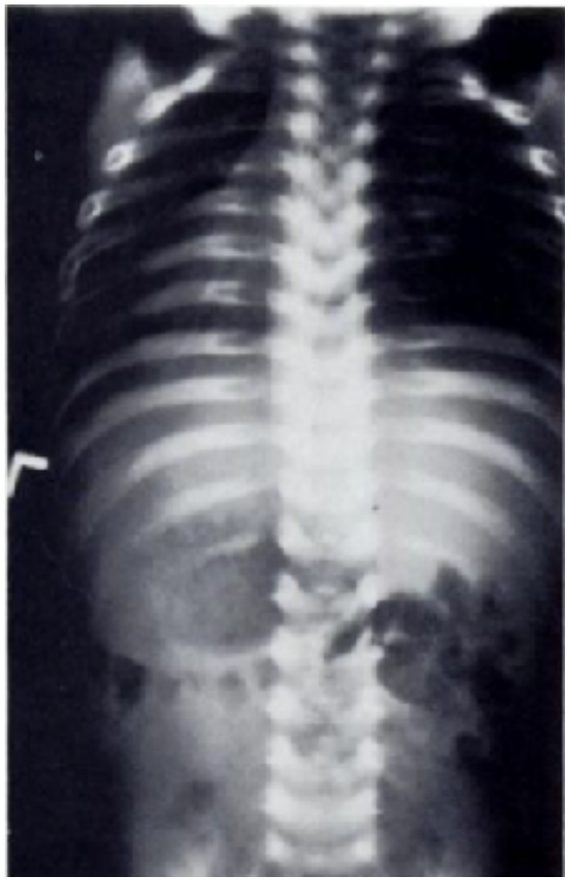
Lenz-Majewski syndrome (LMS)

Radiographic features

- sclerosis of the skull, facial bones, and vertebrae



(Gorlin and Whitley, 1983)



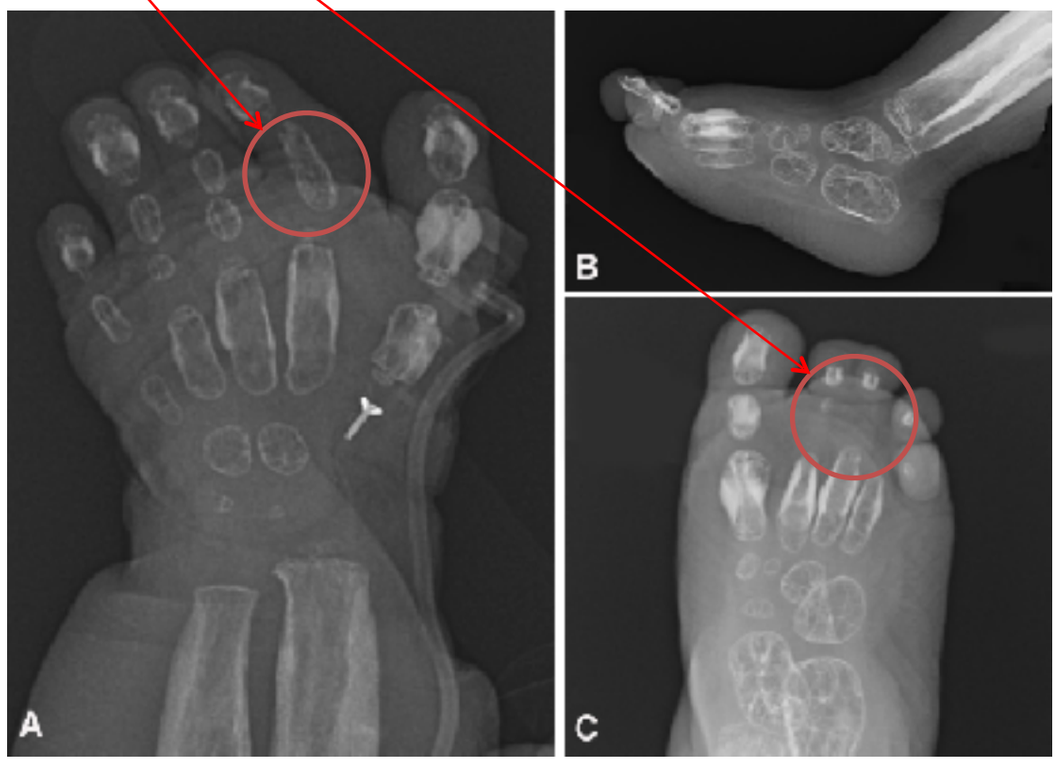
(Gorlin and Whitley, 1983)

- Broad clavicles and ribs
- Diphyseal hyperostosis and undermodeling

Lenz-Majewski syndrome (LMS)

Radiographic features

- Short or absent middle phalanges
- Phalangeal synostosis (fusion of finger bone)
- Abnormal /delayed skeletal maturation



(Shoja *et al.* 2012)

Lenz-Majewski syndrome (LMS)

Etiology

Unknown Genetic Cause

Lenz-Majewski syndrome (LMS)

Mode of Inheritance

no difference in severity or prevalence attributable to the sex

no familial recurrence or consanguinity has been observed

the very specific and consistent phenotype

de novo heterozygous dominant mutations in a single gene

Clinical and radiological manifestations of LMS individuals

➢subject 2 at 7 and 17 years

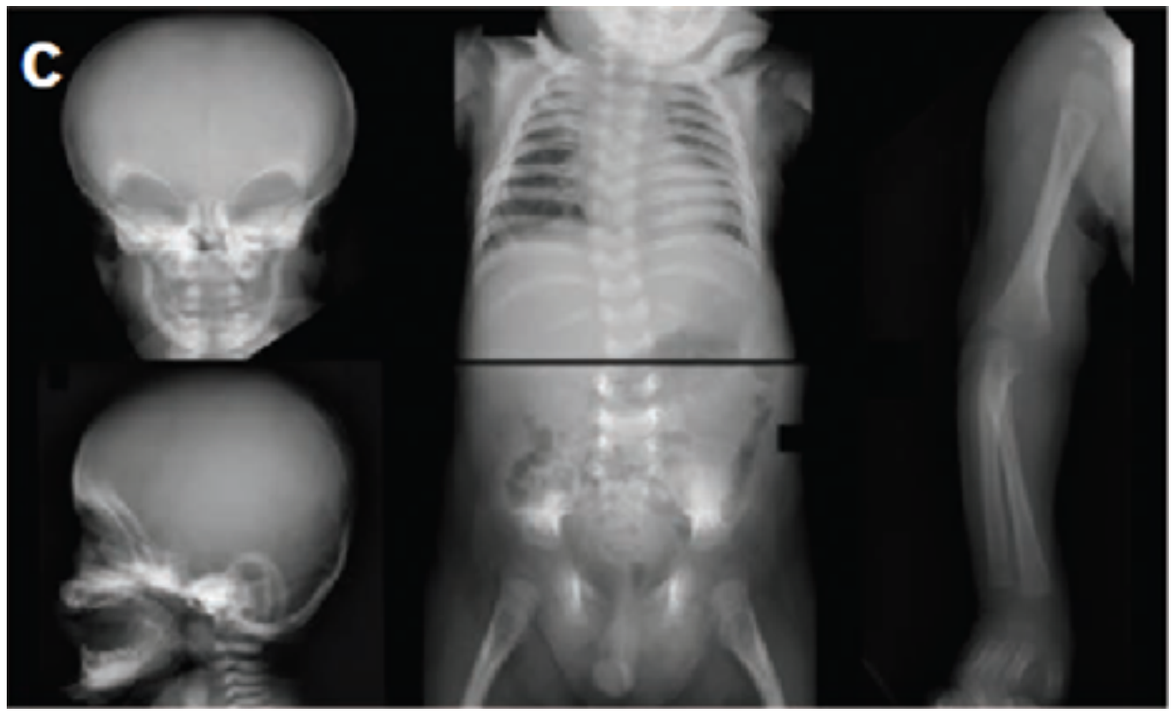
- Sex → female
- Origin → Portuguese
- Current age → 18 years

➢subject 4 at 36 years

- Sex → male
- Origin →Polish
- Current age → 36 years

Clinical and radiological manifestations of LMS individuals

➢subject 1 at 4 months



- Sex → male
- Origin → Kurd – Turkish
- Current age → 2 years

➢subject 3 at 10 years

- Sex → female
- Origin → Thai- Chinese
- Current age → 11.5 years

Clinical and radiological manifestations of LMS individuals

➢subject 5 at 7 years

- Sex → female
- Origin → Czech
- Current age → 9 years

➢**To identify the genetic cause of LMS**

Karyotype analysis

all cases

+

Array comparative genomic hybridization (CGH)

Sub. 1

No pathogenic copy number variants

Whole-exome sequencing

In sub.1-4 + parents of sub.1 and 4

Supplementary Table 3 – Filtering of variants identified by whole-exome sequencing.

Filtering parameters	Variant filtering			
	Pat. 1	Pat. 2	Pat. 3	Pat. 4
Exonic and splice variants[a]	19,488	19,136	18,970	18,344
Heterozygous	11,631	11,431	10,580	10,676
Nonsense, splice or missense variants	5,618	5,580	5,229	5,151
Novel or rare[b]	406	391	449	274
De novo (trio analysis)	35	NA	NA	26
Predicted damaging[c]	19	195	233	13
	Nº of genes with variants in at least x affected individuals			
		x=2	x=3	x=4
Post above filtering		11	4	1

Tips for Handling of Biological Data

1. Understand how data were generated (sources of errors)
2. Check for the errors (quality control)
3. Clarify the research questions (hints on the required tools)
4. Read the manuals! (all the how-to stuff)
5. Understand the context of the research (interpretation of the results)

6. **Stay up-to-date!!**

“A person who never made a mistake never tried anything new.”

–Albert Einstein

Questions?

contact info:

bhoom.suk@mahidol.ac.th

Further Reading

- Thompson & Thompson, Genetics in Medicine, 8e (2016)
- Chap 2: Introduction to the Human Genome
- Chap 3: Gene Structure and Function